Smoothed Spatial Maximum Score Estimation of Spatial Autoregressive Binary Choice Panel Models

Lei, J.

Publication date:
2013

No. 2013-061

SMOOTHED SPATIAL MAXIMUM SCORE ESTIMATION OF SPATIAL AUTOREGRESSIVE BINARY CHOICE PANEL MODELS

By

Jinghua Lei

20 August 2013

ISSN 0924-7815
ISSN 2213-9532

Smoothed spatial maximum score estimation of spatial autoregressive binary choice panel models*

Jinghua Lei †

August 17, 2013

Preliminary and incomplete - Please do not quote or circulate without permission, comments welcome

Abstract

This paper considers spatial autoregressive (SAR) binary choice models in the context of panel data with fixed effects, where the latent dependent variables are spatially correlated. Without imposing any parametric structure of the error terms, this paper proposes a smoothed spatial maximum score (SSMS) estimator which consistently estimates the model parameters up to scale. The identification of parameters is obtained, when the disturbances are time-stationary and the explanatory variables vary enough over time along with an exogenous and time-invariant spatial weight matrix. Consistency and asymptotic distribution of the proposed estimator are also derived in the paper. Finally, a Monte Carlo study indicates that the SSMS estimator performs quite well in finite samples.

JEL classification: C14 C21 C23 C25 R15

Keywords: Spatial Autoregressive Models, Binary Choice, Fixed Effects, Maximum Score Estimation.

*I am very grateful to my advisor Pavel Čížek for his enormously valuable advice, guidance and support. I also benefited from conversations with Zhuojiong Gan, Yufeng Huang, Tobias J. Klein, Lung-Fei Lee, Hong Li, Xiaodong Liu, Geng Niu, Ingmar Prucha, Ruixin Wang, Wendun Wang, Yifang Yu, and thanks for the comments of participants at 7th World Conference of Spatial Econometrics Association at Washington DC. All errors are my responsibility.

†CentER, Department of Econometrics and Operation Research, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Phone: +31-13-466-3399, E-mail: J.Lei@uvt.nl

1 Introduction

The spatial binary choice model has been increasing used in the spatial econometrics literature, especially for the spatial probit model. There are several different specifications of spatial binary choice models, the most popular specification for probit model is the spatial lag probit model: a linear regression model with endogenous interaction effects among the unobserved dependent variable

$$Y_n^* = \lambda_0 W_n Y_n^* + X_n \beta_0 + \epsilon_n, \tag{1}$$

where Y_n^* is an $n \times 1$ vector of latent dependent variable, X_n is an $n \times q$ matrix of independent variables and λ_0 represents the spatial autoregressive coefficient. Endogenous interaction effects are typically considered as the formal specification for the equilibrium outcome of a spatial or social interaction process, in which the value of the dependent variable for one agent is jointly determined with that of neighboring agents. Therefore, this model could also be viewed as the binary choice complete network model with heterogeneous rational expections[1] because usually each agent has a different weight in the network (Lee, Li, and Lin, 2013). In such a model, Y_n^* are interpreted as choice probabilities, which can be derived from the random utility maximization framework McFadden (1974), in the equilibrium under rational expectation. Many studies have considered this model from a methodological viewpoint: McMillen (1992), LeSage (2000), Pace and LeSage (2011) among others. Moreover, Klier and McMillen (2008) replace the probit by the logit specification. Most recently, Qu and Lee (2011) and Jacobs, Samarina, Heijnen, and Elhorst (2013) conduct an important variant of the spatial lag probit model in the following form:

$$Y_n^* = \lambda_0 W_n Y_n + X_n \beta_0 + \epsilon_n,$$

where the latent dependent variable Y_n^* depends on observed choices represented by $W_n Y_n$ rather than unobserved ones. However, one of the basic problems of this interaction model is that the equilibrium may not be unique, so inference is only possible by assuming that one particular equilibrium occurs with probability one over the total number of equilibria.

[1] It is called binary choice complete network model with homogeneous rational expectations when each agent has the same weight (Brock and Durlauf, 2001).

Another specification is a linear regression model with spatially correlated errors:

$$Y_n^* = X_n\beta_0 + v_n, \qquad v_n = \rho_0 W_n v_n + \epsilon_n, \tag{2}$$

where v_n reflects the spatially correlated errors with coefficient ρ_0 and ϵ_n follows a multivariate normal distribution with mean 0 and variance I_n. In this model, the variance of errors is usually normalized to one as it can not be separately identified with the parameter β_0. This spatial error probit model has been studied by Beron and Vijverberg (2004), Fleming (2004), Klier and McMillen (2008) among others. Moreover, Bolduc, Fortin, and Gordon (1997) consider the logit specification in their empirical application such that the probability of $\Pr(y = 1)$ has an analytical solution.

The main assumption of model (1) and (2) is that the distribution of ϵ_n conditional on X is known up to a finite set of parameters, for example, it is often asssumed that ϵ_n has either the normal or logistic distribution. However, when the distribution of ϵ_n is misspecified, estimation methods that require specifying the distribution of ϵ_n yield inconsistent estimators. Furthermore, even if the model is correctly specified, likelihood based estimation methods may suffer from the multidimensional integration problem as the individual error terms are dependent on each other. Many attempts have been proposed to solve this problem, see Jacobs, Samarina, Heijnen, and Elhorst (2013) for a carefully review.

Moreover, estimation would become much more difficult if a context of panel data with fixed effects is considered. Even if the distribution of the errors is correctly specified and there is no spatial dependence, consistently estimating parameters in binary panel models with fixed effect requires clever estimators, such as conditional logit estimation (Chamberlain, 1984) and maximum score estimator (Manski, 1987; Charlier, Melenberg, and van Soest, 1995). These methods could either generate a likelihood function without fixed effects or eliminate the fixed effects by some rank conditions. However, up to my best knowledge, whether these methods still work or not when there is spatial dependence is still unknown. Therefore, the purpose of this paper is to modify the maximum score estimator of Manski (1987) such that it could cnosistently estimate the parameters in spatial lag binary panel models with fixed effects.

In this paper, I consider a fixed effects spatial autoregressive (SAR) binary choice model that is a panel version of model (1), where the only assumption imposed on the errors is time stationary rather than any parametric assumption. Based on this assumption and the exogeneity of time-invariant spatial weight matrix, a similar condition of Lemma 1 in Manski (1987) is derived in this paper. Therefore, a spatial maximum score estimator is defined analogously to that of Manski (1987) and can be smoothed by replacing the sign function with a continuous function as in Horowitz (1992). Although the smoothed spatial maximum score (SSMS) estimator can not be extended to cross-sectional SAR binary choice models, it is applicable to fixed effects SAR binary choice models with arbitrarily spatial correlation in the errors and fixed effects, when such spatial correlation is time-invariant and satisfies some "fading memory" property as described in section 3.2.

The rest of the paper is organized as follows. Section 2 provides the model specification and the suggested smoothed spatial maximum score estimator. Section 3 proves identification, consistency and the asymptotic normality of the proposed estimator. Section 4 presents the results of a Monte Carlo investigation of the finite-sample properties of the estimators and section 5 concludes. All the proofs are provided in the appendices.

2 Spatial Autoregressive Binary Choice Models and SSMS

The SAR binary choice model is

$$Y_{it}^* = \lambda_0 \sum_{j=1}^{n} w_{ij} Y_{jt}^* + X_{it}\beta_0 + \alpha_i + \epsilon_{it}, \qquad i = 1, \ldots, n, t = 1, \ldots, T, \tag{3}$$

where Y_{it}^* is the latent dependent variable that links to the observed binary outcome Y_{it} such that $Y_{it} = 1$ if $Y_{it}^* > 0$ and $Y_{it} = 0$ otherwise, X_{it} are the regressors for the individual i in the t-th period,[2] W_n is a spatial weight matrix with ij-th element w_{ij} and is assumed to be constant across time, λ_0 is a parameter to capture the spatial effect, α_i is the individual fixed effect which is unobserved and allowed to be correlated with the regressors in an arbitrary way, and ϵ_{it} is the idiosyncratic individual error term with a common conditional distribution function, $F_{\epsilon_n}(\cdot, \alpha, X)$ given (α, X) with $X = (X_{n1}, \ldots, X_{nT})$. This spatial

[2]Note that X_{it} consists of time-varying covariates as any time-invariant covariates would be absorbed into the fixed effect α_i.

model is an equilibrium model.

For simplicity and without loss of generality, I consider the case when there are only two time periods (t=1, 2). Suppose that the inverse of matrix $S_n(\lambda_0) = (I_n - \lambda_0 W_n)$ exists and denote $S_n^{-1} = S_n^{-1}(\lambda_0)$, rearrange equation (3) and rewrite it in matrix notation, the equilibrium vector Y_{nt}^* is then

$$Y_{nt}^* = (I_n - \lambda_0 W_n)^{-1}(X_{nt}\beta_0 + \alpha_n + \epsilon_{nt}) = S_n^{-1}(X_{nt}\beta_0 + \alpha_n + \epsilon_{nt}), \quad t = 1, 2. \tag{4}$$

Under the assumption that inverse of matrix $S_n(\lambda_0)$ exists, $S_n^{-1}\epsilon_{nt}$ is a vector of linear combination of the error terms for all individuals. Let e_i denote an $n \times 1$ vector with the i-th element equal to one and all other elements equal to zero, then $e_i^\top S_n^{-1}$ is the i-th row of the matrix $(I_n - \lambda_0 W_n)^{-1}$. Denote $\tilde{\epsilon}_{nt} = e_i^\top S_n^{-1}\epsilon_{nt}$, under the conditional stationarity assumption that ϵ_{n1} and ϵ_{n2} are identically distributed conditional on (α_n, X), we know that $\tilde{\epsilon}_{n1}$ and $\tilde{\epsilon}_{n2}$ also have the same distribution. Therefore, we obtain the following relationship for each individual i as in Lemma 1 of Manski (1987) [3]

$$\begin{aligned}
&E[Y_{i1} - Y_{i2}|\alpha_n, X] > 0 \quad \text{if and only if} \quad e_i^\top S_n^{-1} X_{n1}\beta_0 > e_i^\top S_n^{-1} X_{n2}\beta_0, \\
&E[Y_{i1} - Y_{i2}|\alpha_n, X] = 0 \quad \text{if and only if} \quad e_i^\top S_n^{-1} X_{n1}\beta_0 = e_i^\top S_n^{-1} X_{n2}\beta_0, \\
&E[Y_{i1} - Y_{i2}|\alpha_n, X] < 0 \quad \text{if and only if} \quad e_i^\top S_n^{-1} X_{n1}\beta_0 < e_i^\top S_n^{-1} X_{n2}\beta_0.
\end{aligned} \tag{5}$$

Manski (1987) showed that under some regularity conditions, conditions (5) implies that the true parameter $\theta_0 = (\lambda_0, \beta_0^\top)^\top$ is the unique maximizer of

$$G_i(\theta) = E[\Delta Y_i \operatorname{sgn}\{e_i^\top S_n^{-1}(\lambda)\Delta X_n \beta\}], \quad i = 1, \ldots, n, \tag{6}$$

where $\theta = (\lambda, \beta^\top)^\top$, $\Delta Y_i = (Y_{i1} - Y_{i2})$, $\Delta X_n = X_{n1} - X_{n2}$, $\operatorname{sgn}(x) = 1$ if $x \geq 0$ and -1 otherwise. Apparently, θ_0 is also the unique maximizer of the average of $G_i(\theta)$, that is

$$\theta_0 = \arg\max_\theta \frac{1}{n}\sum_{i=1}^n G_i(\theta) = \arg\max_\theta \frac{1}{n}\sum_{i=1}^n E[\Delta Y_i \operatorname{sgn}\{e_i^\top S_n^{-1}(\lambda)\Delta X_n \beta\}]. \tag{7}$$

[3] A similar result as Corollary of Manski (1987) could also be obtained immediately as

$$M\left(Y_{i1} - Y_{i2}|X, Y_{i1} \neq Y_{i2}\right) = \operatorname{sgn}\{e_i^\top S_n^{-1}(\lambda)\Delta X_n \beta\},$$

where $M\left(Y_{i1} - Y_{i2}|X, Y_{i1} \neq Y_{i2}\right)$ denotes the median of $Y_{i1} - Y_{i2}$ conditional on X and on the event $Y_{i1} \neq Y_{i2}$.

Consistently estimating parameter θ_0 can be obtained by maximizing the following objective function (7):

$$G_n^*(\theta) = \frac{1}{n}\sum_{i=1}^{n} \Delta Y_i \operatorname{sgn}\left(e_i^\top S_n^{-1}(\lambda)\Delta X_n\beta\right).$$

Observe that the behavior of $G_n^*(\cdot)$ is unaffected by removing observations having $Y_{i1} = Y_{i2}$, thus, the estimator maximizing $G_n^*(\cdot)$ is a conditional maximum score estimator. However, it is difficult to derive its asymptotic distribution as the score function is a step function. Chamberlain (1986) has shown that there is no $n^{-1/2}$-consistent estimator of β_0 under Manski's assumptions. Horowitz (1992) then modifies Manski's maximum score estimator (Manski, 1985) by smoothing the score function to be continuous and differentiable, and shows that the convergence rate of the smoothed maximum score estimator is at least as fast as $n^{-2/5}$ and, depending on how smooth the distribution of ϵ_n and $X_n\beta_0$ are, can be arbitrarily close to $n^{-1/2}$. In the context of panel data models with fixed effect, Charlier, Melenberg, and van Soest (1995) investigate the smoothed version of Manski (1987)'s estimator and indicate that maximizing $G_n^*(\theta)$ boils down to maximizing

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\Delta Y_i\left[\operatorname{sgn}\left(e_i^\top S_n^{-1}(\lambda)\Delta X_n\beta\right)+1\right] = \frac{1}{n}\sum_{i=1}^{n}\Delta Y_i \mathbb{1}\{e_i^\top S_n^{-1}(\lambda)\Delta X_n\beta \geq 0\} \quad (8)$$

This objective function can then be smoothed by

$$G_n(\theta;\sigma_n) = \frac{1}{n}\sum_{i=1}^{n}\Delta Y_i K\left(\frac{e_i^\top S_n^{-1}(\lambda)\Delta X_n\beta}{\sigma_n}\right). \quad (9)$$

where $K(u)$ is some smooth function that converges to the indicator function as $n \to \infty$, and σ_n is a sequence of strictly positive real numbers satisfying $\lim_{n\to\infty}\sigma_n = 0$. Note that $K(u)$ could be a cumulative distribution function such as the cumulative standard normal distribution function $\Phi(u)$ as in Horowitz (1992), and σ_n can be viewed as the bandwidth.

Remark 1. Apparently, when there is no spatial effect ($\lambda_0 = 0$), then the SSMS estimator that maximizes equation (9) degenerates to the standard smoothed maximum score estimator for panel models as in Charlier, Melenberg, and van Soest (1995).

Moreover, the identification and estimation strategy described above also works for models with arbitrarily spatially correlated errors and fixed effects, if the spatial correlation

is time stationary and satisfies the "fading memory" property as stated in the next section. For example, the mixed spatial lag and spatial error binary choices models with fixed effects:

$$Y_{nt}^{*} = \lambda_0 W_{n,1} Y_{nt}^{*} + X_{nt}\beta_0 + W_{n,3}\alpha_n + v_{nt}, \qquad v_{nt} = \rho_0 W_{n,2} v_{nt} + \epsilon_{nt}.$$

Suppose that the inverse of matrices $S_n(\lambda_0) = (I - \lambda_0 W_{n,1})$, $(I - \rho_0 W_{n,2})$ exists and the spatial weighting matrices $W_{n,1}$, $W_{n,2}$ and $W_{n,3}$ are time-invariant, rearranging the above equation and rewrite it in matrix notation, we have

$$\begin{aligned} Y_{nt}^{*} &= (I_n - \lambda_0 W_{n,1})^{-1} \left[X_{nt}\beta_0 + W_{n,3}\alpha_n + (I_n - \rho_0 W_{n,2})^{-1}\epsilon_{nt} \right] \\ &= S_n^{-1}(X_{nt}\beta_0 + W_{n,3}\alpha_n) + S_n^{-1}(I_n - \rho_0 W_{n,2})^{-1}\epsilon_{nt}. \end{aligned}$$

As in equation (5), when there are only two time periods (t=1, 2) and ϵ_{n1} and ϵ_{n2} are identically distributed conditional on (α_n, X), then $\bar{\epsilon}_{n1}$ and $\bar{\epsilon}_{n2}$ also have the same distribution, where $\bar{\epsilon}_{nt} = [S_n^{-1}(I_n - \rho_0 W_{n,2})^{-1}]_i \epsilon_{nt}$ and $e_i^{\top} S_n^{-1}(I - \rho_0 W_{n,2})^{-1}$ denotes the i-th row of the matrix $(I - \lambda_0 W_{n,1})^{-1}(I - \rho_0 W_{n,2})^{-1}$. Therefore, we could also obtain the same relationship for each individual i as in equation (5). The identification and estimation strategy will be exactly the same as I dicussed previously, however, the parameter ρ_0 can not be estimated in this case.

In addition, when $\beta_0 = 0$, the model degenerates to a spatial binary choice models without covariates. In this case, the spatial effect λ_0 is not identified without imposing additional assumptions on the error terms. Another point that we should notice is that the identification strategy described in this paper cannot be applied to the cross sectional spatial binary choice models directly.

Finally, when the time periods are more than two but finite, the SSMS estimator can be defined analogously to that of Charlier, Melenberg, and van Soest (1995) by

$$\widehat{\theta}_{nT} = \arg\max_{\theta} \frac{1}{nT(T-1)} \sum_{i=1}^{n} \sum_{s<t} c_{its}(Y_{it} - Y_{is}) K\left(\frac{e_i^{\top} S_n^{-1}(\lambda)(X_{nt} - X_{ns})\beta}{\sigma_n} \right),$$

where $c_{its} = r_{it} r_{is}$, with $r_{it} = 1$ if $\{Y_{it}, X_{it}\}$ is observed, and zero otherwise. Therefore, $c_{its} = 1$ if both $\{Y_{it}, X_{it}\}$ and $\{Y_{is}, X_{is}\}$ are observed and zero otherwise. The inclusion of c_{its} is to make the SSMS estimator be applicable to an unbalanced panel, which is common

in applications.

Remark 2. First we note that the objective function of (9) is identical to the absolute loss objective function

$$\widehat{\theta} = \arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \left| \Delta Y_i - K\left(\frac{e_i^\top S_n^{-1}(\lambda) \Delta X_n \beta}{\sigma_n} \right) \right|,$$

and the squared loss objective function

$$\widehat{\theta} = \arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \left[\Delta Y_i - K\left(\frac{e_i^\top S_n^{-1}(\lambda) \Delta X_n \beta}{\sigma_n} \right) \right]^2.$$

Khan (2012) proposes that when a standard normal distribution $\Phi(\cdot)$ is applied for the kernel function, then we can define the spatial nonlinear least square (SNLLS) probit estimator as

$$\widehat{\theta}_{SNLLS} = \arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \left[\Delta Y_i - \Phi\left(\frac{e_i^\top S_n^{-1}(\lambda) \Delta X_n \beta}{\sigma_n} \right) \right]^2.$$

The main advantage of this procedure is that the standard NLLS objective function can be extended to the case with spatial correlation, and the standard software packages, such as Stata, can be easily adjusted to compute the SNLLS estimator.

3 Identification and asymptotic properties of SSMS

3.1 Identification

In this subsection, identification of parameters in model (3) is provided, and the definition of identication is similar as in Manski (1987). Consider $(\lambda, \beta^\top)^\top \in \Lambda \times R^q, (\lambda, \beta^\top)^\top \neq (\lambda_0, \beta_0^\top)^\top$, conditions (5) distinguishes $(\lambda, \beta^\top)^\top$ from $(\lambda_0, \beta_0^\top)^\top$ if there exists a set of ΔX values having positive $F_{\Delta X}$ probability such that condition (5) does not hold when $(\lambda, \beta^\top)^\top$ is substituted for $(\lambda_0, \beta_0^\top)^\top$. In this case, let

$$V_{(\lambda,\beta)} = \left[\Delta X \in R^q : \operatorname{sgn}(e_i^\top S_n^{-1}(\lambda) \Delta X_n \beta) \neq \operatorname{sgn}(e_i^\top S_n^{-1} \Delta X_n \beta_0) \right] \tag{10}$$

we say that $(\lambda_0, \beta_0^\top)^\top$ is identified relative to $(\lambda, \beta^\top)^\top$ if

$$R(\lambda, \beta) \equiv \int_{V_{(\lambda,\beta)}} dF_{\Delta X} > 0 \tag{11}$$

Assumption 1. *i). $F_{\epsilon_{i1}|X,\alpha_n} = F_{\epsilon_{i2}|X,\alpha_n}$ for all i and (X, α_n).*

ii). The support of $F_{\epsilon_{i1}|X,\alpha_n}$ is $\mathbb{R}^1$ for all i and (X, α_n).

Assumption 2. *i). The support of $F_{\Delta X}$ is not contained in any proper linear subspace of $\mathbb{R}^q$.*

ii). There exists at least one $q' \in [1, 2, \ldots, q]$ such that $\beta_{0,q'} \neq 0$, and for almost every value of $\Delta\tilde{X}_i = (\Delta X_{i,1}, \Delta X_{i,2}, \ldots, \Delta X_{i,q'-1}, \Delta X_{i,q'+1}, \ldots, \Delta X_{i,q})^\top$, the scalar random variable $\Delta X_{i,q'}$ has everywhere positive Lebesgue density conditional on $\Delta\tilde{X}_i$ for all $i = 1, 2, \ldots, n$ and conditional on $\Delta X_{j,q'}$ for all $j \neq i$.

Assumption 3. *The matrix $S_n(\lambda_0) = I_n - \lambda_0 W_n$ is nonsingular;*

Assumptions 1 and 2 have the same forms as Assumptions 1 and 2 in Manski (1987), except that we have different conditionings. As in assumption 1 i), ϵ_{it} is stationary not only conditional on its own characteristics, but also conditional on the characteristics of other members. Such conditioning also appears in Assumption 2, and is necessary because there is spatial correlation between individuals, which is obvious if we assume the process $\{X_{it}, \alpha_i, \epsilon_{it}\}$ is strong mixing as in section 3.2. Assumption 1 ii) guarantees that the event $Y_{i1} \neq Y_{i2}$ occurs with positive probability for all α_n. Assumption 2 i) is the familiar full-rank condition that prevents a global failure of identification, and part ii) is a substantive restriction, which implies that $\Delta X_n\beta$ has everywhere positive density for all β such that $\beta_{q'} \neq 0$. Assumption 3 guarantees that the system (3) has an equilibrium and matrix $S_n(\lambda_0)$ is invertible.

Clearly, the scale of β_0 is not identified. To see this, we can simply set $\lambda = \lambda_0$, then the identification problem degenerates to that of Manski (1987). Identification of θ_0 requires that there is a positive probability such that $e_i^\top S_n^{-1}(\lambda)\Delta X_n\beta$ has a different sign with $e_i^\top S_n^{-1}\Delta X_n\beta_0$. As Assumption 2 imposes no condition on the parameter vector θ_0 except that $\beta_{0,q'} \neq 0$, it is possible for $e_i^\top S_n^{-1}(\lambda)\Delta X_n\beta$ to have bounded support for all θ, given sharper bounds on θ_0. Therefore, θ_0 is identified, which is stated in the following Lemma.

Lemma 1. Under Assumptions 1-3, $(\lambda_0, \beta_0^\top)^\top$ is identified relative to $(\lambda, \beta^\top)^\top \in \Lambda \times R^q$, where $\beta/||\beta|| \neq \beta_0/||\beta_0||$.

3.2 Consistency

In this subsection, consistency of estimators that maximize the objective function (9) is established. The main difficulty to prove the consistency is that the objective function (9) is based on a dependent and heterogenous process. Therefore, some appropriate "fading memory" property must be guaranteed to support laws of large numbers and uniform laws of large numbers, and the "fading memory" property that I will prove is near epoch dependence which is defined in Definition 1.

To proceed, we need to first define the space and metric (which are not restricted to physical space and distance) for the convenience of analyzing the spatial correlation structure. Following Jenish and Prucha (2009, 2012), I consider spatial processes located on a (possibly) unevenly spaced lattice that satisfies the following assumption.

Assumption 4. *The lattice $D \subseteq \mathbb{R}^d, d \geq 1$ is infinite countable. All elements in D are located at distances of at least $d_0 > 0$ from each other, i.e., for all $l_i, l_j \in D : d(l_i, l_j) \geq d_0$, where l_i denotes some location of corresponding unit i; without loss of generality, we assume that $d_0 = 1$.*

The assumption of a minimum distance ensures the growth of the sample size as the sample regions $D_n = \{l_1, \ldots, l_n\} \subset D$ expand, which means the asymptotic methods that I employ in this paper are increasing domain asymptotics.

The models considered in this paper are actually the Cliff and Ord (1981) type models, which is one of the common aproaches to model cross-sectional dependence in the econometrics literature. In the Cliff-Ord type models, the spatial weights $w_{ij,n}$ depend on some measure of distance d_{ij} and decline as the distance increases. Under Assumption 3, model (4) is then $Y_{it} = \mathbb{1}\{e_i^\top S_n^{-1}[X_{nt}\beta_0 + \alpha_n + \epsilon_{nt}] > 0\}$, where $e_i^\top S_n^{-1}$ could be denoted by a vector $(a_{i1}, \ldots, a_{in})$. Although for fixed n, the output process Y_{it} only depends on a finite number of elements of the input process $\eta_{it} = (X_{it}, \alpha_i, \epsilon_{it})^\top$, the mixing property of η_{it} may not carry over to Y_{it}. The reason is that the number of elements composing the spatial lags grows unboundedly with the sample size so that the mixing property can break down

in the limit. This is especially important when analyzing the asymptotic properties of Cliff-Ord type processes. Therefore, towards establishing that $\{Y_{it}, l_i \in D_n\}$ is near epoch dependent (NED) on $\{\eta_{it}, l_i \in D_n\}$, we maintain the following assumptions:

$$\lim_{d\to\infty} \sup_n \sup_{1\leq i\leq n} \sum_{1\leq j\leq n: d(l_i,l_j)>d} |a_{ij}| = 0 \tag{12}$$

and

$$\sup_n \sup_{1\leq i\leq n,t} ||\eta_{it}||_p < \infty \quad \text{for some} \quad p \geq 1. \tag{13}$$

Jenish and Prucha (2012) show that a sufficient condition for (12) is that for some $\gamma > 0$,

$$\sup_n \sup_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| d(l_i, l_j)^\gamma < \infty.$$

A similar condition has been used recently by Kelejian and Prucha (2007), and should be satisfied in a wide range of applications. It is slightly stronger than the typical assumption in the Cliff-Ord literature which imposes that the row and column sums of the absolute elements of the matrix S_n^{-1} are uniformly bounded as in Assumption 6 ii).

Now I define the near epoch denpendence (NED) of random variables Y_{it} based on a process of random variables η_{it} as follows:

Definition 1. Random variables Y_{it}, $i = 1, \ldots, n$, $t = 1, 2$ are called near epoch dependent on η_{it} if

$$\sup_i ||Y_{it} - E(Y_{it}|\Im_{i,n}(m))||_2 = d_t \nu(m) \to 0, \quad \text{as} \quad m \to \infty \tag{14}$$

where d_i is a sequence of positive constant and $\Im_{i,n}(m) = \sigma\left(\eta_{jt,n} : d(l_i, l_j) \leq m\right)$ is the σ field generated by the random variables $\eta_{jt,n}$ located in the m-neighborhood of location i.

The idea behind the near epoch dependence condition is that given the m-neighborhood of imput variables η_{it}, Y_{it} should be predictable up to arbitrary accuracy. That is, the approximation error declines "sufficiently fast" as the conditioning set of input variables expands. The base process η_{it} needs to satisfy a condition such as strong or uniform mixing or independence.

Assumption 5. $\{\eta_{it}\}, i = 1, \ldots, n, t = 1, 2$, *is a strict stationary strong mixing process with* α*-mixing coefficient* $\alpha(m)$.

Proposition 1. Under Assumptions 1, 3-5 and conditions (12)-(13), the process $\{Y_{it}\}$, $\{\Delta Y_i\}$, and $\{\text{sgn}(e_i^\top S_n^{-1}(\lambda)\Delta X_n\beta)\}$ are uniformly NED on the process $\{\eta_{it}\}$.

Remark 3. Proposition 1 shows that Y_{it} is a sequence of 0/1 valued random variables that is near epoch dependent on η_{it}. Then (Y_{it}, η_{it}) is strong mixing by Theorem 2 of de Jong and Woutersen (2011), and the mixing property of (Y_{it}, η_{it}) will be used in the proofs for consistency and asymptotic normality of the smoothed spatial maximum score estimator. Although $\{Y_{it}\}$ is strong mixing, it is not stationary as the inverse spatial weights $e_i^\top S_n^{-1}(\lambda)$ are different for each individual i in general. One example for $\{Y_{it}\}$ to be stationary is where the spaital correlation only exists within groups of the same size, and equal weights are assigned for individuals in the same group.

Assumption 6. *i).* $|\beta_{0,q}| = 1$ *and* $\tilde{\beta}_0 = (\beta_{0,1}, \ldots, \beta_{0,q-1})^\top$ *is contained in a compact subset* $\mathcal{B}$ *of* $\mathbb{R}^{q-1}$*;*

ii). The sequence $\{W_n\}$ *and* $\{S_n^{-1}\}$ *are uniformly bounded in both row and column sums;* [4]

iii). $\{S_n^{-1}(\lambda)\}$ *are uniformly bounded in either row or column sums, uniformly in* λ *in a compact parameter space* Λ*. The true parameter* λ_0 *is in the interior of* Λ.

The uniform boundedness condition of S_n^{-1} in Assumption 6 ii) implies that $S_n^{-1}(\lambda)$ are uniformly bounded in both row and column sums uniformly in a neighborhood of λ_0 (Lee, 2004). Assumption 6 i) and iii) are needed to deal with the nonlinearity of $K\left(e_i^\top S_n^{-1}(\lambda)\Delta X_n\beta/\sigma_n\right)$ as a function of λ and β. The parameter space $\Lambda \times R^q$ is usually assumed to be a compact convex subset of R^{q+1} for a nonlinear extremum estimation, this assumption is required for the uniform convergence of the sample average objective function in the proof of consistency (Amemiya, 1985). However, Wang and Lee (2012)

[4] The notions of uniform boundedness can be defined in terms of some matrix norms: the maximum column matrix norm $||\cdot||_1$ of a $n \times n$ matrix $A = (a_{ij})$ is defined as $||A_n||_1 = max_{1\le j\le n}\sum_{i=1}^n |a_{ij}|$, and the maximum row sum matrix norm $||\cdot||_\infty$ is $||A_n||_\infty = max_{1\le i\le n}\sum_{j=1}^n |a_{ij}|$ (see Horn and Johnson (1985), pp.294-295). The uniformly boundedness of $\{A_n\}$ in column (resp. row) sums is equivalent to the sequence $\{||A_n||_1\}$ (resp. $\{||A_n||_\infty\}$) being bounded.

Lemma A.2 of Lee (2004) shows that, for any weights matrix, $||\lambda_0 W_n||_1 < 1$ and $||\lambda_0 W_n||_\infty < 1$ for all n, are sufficient conditions for S_n^{-1} to be uniformly bounded in both row and column sums.

Because a matrix norm $||\cdot||$ has the submultiplicative property that $||A_nB_n|| \le ||A_n||\cdot||B_n||$, Assumption 6 guarantees that products of matrices in our analysis such as $S_n^{-1}W_nS_n^{-1}$ and $S_n^{-1}W_nS_n^{-1}W_nS_n^{-1}$, etc., will be uniformly bounded in row and column sums.

mention that relaxation of this assumption would be an important issue of future research as it does not cover leading specification for the parameter space of λ, which is often taken to be an open set, e.g.,$(-1, 1)$.

Under Assumptions 1-6, the following theorem shows the consistency of the smoothed spatial maximum score estimator.

Theorem 1. *Let Assmptions 1-6 hold. Let θ_n be a solution to*

$$\max_{\theta} G_n(\theta; \sigma_n), \tag{15}$$

where $G_n(\theta; \sigma_n) = \frac{1}{n}\sum_{i=1}^{n} \Delta Y_i K\left(\frac{e_i^\top S_n^{-1}(\lambda)\Delta X_n \beta}{\sigma_n}\right)$ and $\theta = (\lambda, \beta^\top)^\top$, then $\theta_n \to^p \theta_0$. If in addition the strong mixing coefficients satisfy $\alpha(m) \leq Cm^{-r}$ for positive constants C and r, then $\theta_n \to^{a.s.} \theta_0$.

3.3 Asymptotic normality

In this subsection, the asymptotic normality of the smoothed spatial maximum score estimator is establised, and the approach is analogous to that of Horowitz (1992) and de Jong and Woutersen (2011) except that the asymptotic properties are built on a dependent and heterogenous process while the process in Horowitz (1992) is i.i.d and the process in de Jong and Woutersen (2011) is dependent but stationary.

Let Assumptions $1-3$ hold, and suppose $K(\cdot)$ is twice differentiable everywhere. Then $G_n(\theta; \sigma_n)$ is twice differentiable with respect to $\tilde{\theta} = (\lambda, \tilde{\beta}^\top)^\top$, where $\tilde{\beta} = (\beta_1, \ldots, \beta_{q-1})^\top$. Assumption 6 ensures that $\tilde{\theta}_0$ is an interior point of $\tilde{\Theta}$. Define $T_n(\theta; \sigma_n) = \partial G_n(\theta; \sigma_n)/\partial\tilde{\theta}$, and $Q_n(\theta; \sigma_n) = \partial^2 G_n(\theta; \sigma_n)/\partial\tilde{\theta}\partial\tilde{\theta}^\top$. Let $\theta_n \equiv (\tilde{\theta}_n^\top, \beta_{n,q})^\top$ denote a solution to problem (15), then with probability approaching 1 as $n \to \infty$, $\tilde{\theta}_n$ is an interior point of $\tilde{\Theta}$, $\beta_{n,q} = \beta_{0,q} = \pm 1$, and $T_n(\theta_n; \sigma_n) = 0$. A Taylor series expansion of $T_n(\theta_n; \sigma_n)$ yields

$$T_n(\theta_n; \sigma_n) = T_n(\theta_0; \sigma_n) + Q_n(\theta_n^*; \sigma_n)(\tilde{\theta}_n - \tilde{\theta}_0) = 0, \tag{16}$$

where θ_n^* is between θ_n and θ_0. Similar to Horowitz (1992), if there is a real function $\rho(n)$ such that $\rho(n)T_n(\theta_0; \sigma_n)$ converges in distribution as $n \to \infty$, and suppose $Q_n(\theta_n^*; \sigma_n)$

converges in probability to a nonsingular and nonstochastic matrix Q. Then

$$\rho(n)(\tilde{\theta}_n - \tilde{\theta}_0) = -Q^{-1}\rho(n)T_n(\theta_0;\sigma_n) + o_p(1). \tag{17}$$

Thus, we know that $\tilde{\theta}_n - \tilde{\theta}_0$ converges to 0 at the rate of $\rho(n)^{-1}$ and $\rho(n)(\tilde{\theta}_n - \tilde{\theta}_0)$ is distributed asymptotically as $-Q^{-1}\rho(n)T_n(\theta_0;\sigma_n)$.

Let $z_i = e_i^\top S_n^{-1}\Delta X_n\beta_0 = e_i^\top S_n^{-1}\Delta\tilde{X}_n\tilde{\beta}_0 + e_i^\top S_n^{-1}\Delta X_{n,q}$, then there is a one-to-one relation between $\left(\Delta\tilde{X}_n, Z_n\right)$ and ΔX_n for any fixed θ_0, where $Z_n = (z_1,\ldots,z_n)^\top$. Denote $Z_{-i} = (z_1,\ldots,z_{i-1},z_{i+1},\ldots,z_n)^\top$ and $\tilde{Z}_i = \{\Delta\tilde{X}_n, Z_{-i}\}$. By Assumption 2, the distribution of z_i conditional on $\tilde{Z}_i$ has everywhere positive density with respect to Lebesgue measure for almost every $\tilde{Z}_i$ and $i = 1,\ldots n$. Let $p_i(z_i|\tilde{Z}_i)$ denote this density. For each positive integer j, define $p_i^{(j)}(z_i|\tilde{Z}_i) = \partial^j p_i(z_i|\tilde{Z}_i)/\partial z_i^j$ whenever the derivative exists, and define $p_i^{(0)}(z_i|\tilde{Z}_i) = p_i(z_i|\tilde{Z}_i)$. Let P_i denote the cumulative distribution function of $\tilde{Z}_i$, and let $F_i(\cdot|z_i,\tilde{Z}_i)$ denote the cumulative distribution of $\tilde{\epsilon_i} = e_i^\top S_n^{-1}(\epsilon_{n1} - \epsilon_{n2})$ conditional on z_i and $\tilde{Z}_i$. For each positive integer j, define $F_i^{(j)}(-z_i|z_i,\tilde{Z}_i) = \partial^j F_i^{(j)}(-z_i|z_i,\tilde{Z}_i)/\partial z_i^j$ whenever the derivative exists. Define the scalar constants α_A and α_D by $\alpha_A = \int_{-\infty}^{\infty} v^h K'(v)dv$ and $\alpha_D = \int_{-\infty}^{\infty}[K'(v)]^2 dv$ whenever these quantities exist. For each integer $h \geq 2$, define the $q \times 1$ vecter A and the $q \times q$ matrices D and Q by

$$A = -2\alpha_A \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{h}\left(\frac{1}{k!(h-k)!}E\left[F_i^{(k)}(0|0,\tilde{Z}_i)p_i^{(k)}(z_i|\tilde{Z}_i)\tilde{B}_{1,i}\right]\Pr(Y_{i1}\neq Y_{i2})\right),$$

$$D = \alpha_D \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}E\left[p_i(0|\tilde{Z}_i)\tilde{B}_{1,i}\tilde{B}_{1,i}^\top\right]\Pr(Y_{i1}\neq Y_{i2}),$$

$$Q = \lim_{n\to\infty}\frac{2}{n}\sum_{i=1}^{n}E\left[F_i^{(1)}(0|0,\tilde{Z}_i)p_i(0|\tilde{Z}_i)\tilde{B}_{2,i}\right]\Pr(Y_{i1}\neq Y_{i2}),$$

where $\tilde{B}_{1,i} = \left(e_i^\top S_n^{-1}W_n S_n^{-1}\Delta X_n\beta_0, e_i^\top S_n^{-1}\Delta\tilde{X}_n\right)^\top$ and

$$\tilde{B}_{2,i} = \begin{pmatrix} \left(e_i^\top S_n^{-1}W_n S_n^{-1}\Delta X_n\beta_0\right)^2 & e_i^\top S_n^{-1}W_n S_n^{-1}\Delta X_n\beta_0 e_i^\top S_n^{-1}\Delta\tilde{X}_n \\ * & \Delta\tilde{X}_n^\top[S_n^{-1}]^\top e_i e_i^\top S_n^{-1}\Delta\tilde{X}_n \end{pmatrix}.$$

Assumption 7. *i) The α-mixing coefficient satisfies $\alpha(m) \leq Cm^{-(2s-2)/(s-2)-\gamma}$ for some*

$\gamma > 0$.

ii) For some sequence $m_n \geq 1$ *and when* $n \to \infty$,

$$\sigma_n^{-3(p+q-1)}\sigma_n^{-2}n^{1/s}\alpha(m_n)+\sigma_n^{-2(p+q-1)/\mu}n^{2/s}\alpha(m_n)+|\log(nm_n)|\left(n^{1-4/s}\sigma_n^{-4}m_n^{-2}\right)^{-1} \to 0.$$

Assumption 8. *For all vectors* ξ *such that* $|\xi| = 1$, $E|\xi^\top \tilde{B}_{1,i}|^s < \infty$ *for some* $s > 4$ *and all* i.

These two assumptions are identical to Assumptions 6 and 7 of de Jong and Woutersen (2011). They strengthen the fading memory conditions of Assumption 5 in order to establish asymptotic normality.

The following assumptions are analogous to Assumptions 7-11 of Horowitz (1992):

Assumption 9. *i)* $K(\cdot)$ *is twice differentiable everywhere,* $|K'(\cdot)|$ *and* $|K''(\cdot)|$ *are uniformly bounded, and each of the following integrals over* $(-\infty, \infty)$ *is finite:* $\int [K'(v)]^4 dv$, $\int [K''(v)]^2 dv$, $\int |v^2 K''(v)| dv$.

ii) For some integer $h \geq 2$ *and each integer* $k (1 \leq k \leq h)$, $\int |v^k K'(v)| dv < \infty$, *and*

$$\int_{-\infty}^{\infty} v^k K'(v) dv = \begin{cases} 0 & \text{if } k < h, \\ d & \text{(nonzero) if } k = h. \end{cases}$$

iii) For any integer k *between 0 and* h, *any* $\gamma > 0$, *and any sequence* $\{\sigma_n\}$ *converging to 0,*

$$\lim_{n\to\infty} \sigma_n^{k-h} \int_{|\sigma_n v|>\gamma} |v^k K'(v)| dv = 0, \quad \lim_{n\to\infty} \sigma_n^{-1} \int_{|\sigma_n v|>\gamma} |K''(v)| dv = 0.$$

Assumption 10. *For all* i *and each integer* k *such that* $1 \leq k \leq h-1$, *all* z_i *in a neighborhood of 0, almost every* $\tilde{Z}_i$, *and some* $M < \infty$, $p_i^{(k)}(z_i|\tilde{Z}_i)$ *exists and is a continuous function of* z_i *satisfying* $p_i^{(k)}(z_i|\tilde{Z}_i) < M$. *In addition,* $|p_i(z_i|\tilde{Z}_i)| < M$ *for all* z_i *and almost every* $\tilde{Z}_i$.

Assumption 11. *For all* i *and each integer* k *such that* $1 \leq k \leq h$, *all* z_i *in a neighborhood of 0, almost every* $\tilde{Z}_i$, *and some* $M < \infty$, $F_i^{(k)}(-z_i|z_i, \tilde{Z}_i)$ *exists and is a continuous function of* z_i *satisfying* $F_i^{(k)}(-z_i|z_i, \tilde{Z}_i) < M$.

Assumption 12. *The true parameter* $\tilde{\theta}_0$ *is an interior point of* $\tilde{\Theta}$.

Assumption 13. *The matrix Q is negative definite.*

In addition to the above assumptions, we still need the following two assumptions that similar to Assumptions 13 and 14 in de Jong and Woutersen (2011). The first assumption is needed to ensure proper behavior of covariance terms, and the second assumption on $K''(\cdot)$ is needed to formally show a uniform law of large numbers for the second derivative of the objective function.

Assumption 14. *The conditional joint density $p(z_i, z_j|\tilde{Z}_i, \tilde{Z}_j)$ exists and is continuous at $(z_i, z_j) = (0, 0)$ for all $i \neq j$.*

Assumption 15. *$K''(\cdot)$ satisfies, for some $\mu \in (0, 1]$ and $L \in [0, \infty)$ and all $x, y \in \mathbb{R}$,*

$$|K''(x) - K''(y)| \leq L|x - y|^{\mu}.$$

The main results concerning the asymptotic distribution of the smoothed spatial maximum score estimator are given by the following theorem.

Theorem 2. *Let Assumptions 1-15 hold for some $h \geq 2$, then*
(a) If $n\sigma_n^{2h+1} \to \infty$ as $n \to \infty$, $\sigma_n^{-h}(\tilde{\theta}_n - \tilde{\theta}_0) \to^p -Q^{-1}A$.
(b) If $n\sigma_n^{2h+1} \to \infty$ has a finite limit κ as $n \to \infty$, then

$$\sigma_n^{-h}(\tilde{\theta}_n - \tilde{\theta}_0) \to^d N(-\kappa^{1/2}Q^{-1}A, \quad Q^{-1}DQ^{-1}).$$

In order to make the results of Theorem 2 useful in applications, the next theorem shows how A, D and Q could be consistently estimated from observations of (Y_{nt}, X_{nt}, W_n).

Theorem 3. *Let θ_n be a consistent smoothed spatial maximum score estimator based on σ_n such that $\sigma_n \propto n^{-1/(2h+1)}$. For $\theta \in \{-1, 1\} \times \tilde{\Theta}$, define*

$$t_i(\theta, \sigma) = \mathbb{1}\{Y_{i1} \neq Y_{i2}\}\,(2 \cdot \mathbb{1}\{Y_{i1} = 1, Y_{i2} = 0\} - 1)\,B_i^{(1)}(\theta, \sigma),$$

where $B_i^{(1)}(\theta, \sigma)$ is defined in Appendix A. Let σ_n^* be such that $\sigma_n^* \propto n^{-\delta/(2h+1)}$, where $0 < \delta < 1$. Then: (a) $\widehat{A}_n = (\sigma_n^*)^{-h}T_n(\theta_n, \sigma_n^*)$ converges in probability to A; (b) the matrix

$$\widehat{D}_n \equiv \frac{\sigma_n}{n}\sum_{i=1}^{n} t_i(\theta_n, \sigma_n)t_i(\theta_n, \sigma_n)^{\top}$$

converges in probability to D; (c) $Q_n(\theta_n, \sigma_n)$ converges in probability to Q.

Theorem 2 indicates that the asymptotic bias of $n^{h/(2h+1)}(\tilde{\theta}_n - \tilde{\theta}_0)$ is $-\kappa^{h/(2h+1)}Q^{-1}A$ if $\sigma_n \propto n^{-1/(2h+1)}$, and this can be consistently estimated by $-\kappa^{h/(2h+1)}Q_n(\theta_n, \sigma_n)^{-1}\widehat{A}_n$ by Theorem 3. Therefore, an asymptotically unbiased estimator of $\tilde{\theta}_0$, which is also called the bias-corrected smoothed spatial maximum score estimator, is

$$\widehat{\theta}_{bc} = \tilde{\theta}_n + (\kappa/n)^{h/2h+1}Q_n(\theta_n, \sigma_n)^{-1}\widehat{A}_n. \tag{18}$$

Another important issue in applications is choosing the bandwidth σ_n, and no completely satisfactory solutions have been found for the well-known problem of bandwidth selection. Horowitz (1992) proposed that a possible choice of the bandwidth for smoothed maximum score estimator is $(\hat{\kappa}/n)^{1/(2h+1)}$, where $\hat{\kappa}$ is a consistent estimate of κ^*. κ^* is the asymptotically optimal value of κ that minimizing MSE of the smoothed maximum score estimator, as shown in part (c) of Theorem 2 in Horowitz (1992), $\kappa^* = [\text{trace}(Q^{-1}\Omega Q^{-1}D)]/(2hA^\top Q^{-1}\Omega Q^{-1}A)$ for any nonstochastic, positive semidefinite matrix such that $A^\top Q^{-1}\Omega Q^{-1}A \neq 0$. Therefore, the procedure of bandwidth selection is as follows. Given h, first choose any $\sigma_n \propto n^{-1/(2h+1)}$ to compute the smoothed maximum score estimate θ_n, then use θ_n and any $\sigma_n^* \propto n^{-\delta/(2h+1)}, (0 < \delta < 1)$ to compute $\widehat{A}_n, \widehat{D}_n$, and $Q_n(\theta_n, \sigma_n)$. After that, estimate κ^* from the formule given above by replacing A, D, and Q with $\widehat{A}_n, \widehat{D}_n$, and $Q_n(\theta_n, \sigma_n)$. Finally, the bandwidth is given by $(\kappa^*/n)^{1/(2h+1)}$.

In finite samples, $E\widehat{A}_n \neq A$. The bias of $\widehat{A}_n$ consists of two components: one component is due to the use of a nonzero bandwidth to estimate A, and the other is due to the use of an estimate of θ_0 in the estimator of A. As suggested in Horowitz (1992), only the second component of the bias can be removed by a corrected estimator of A, which is given by

$$\widehat{A}_n^* = \frac{\widehat{A}_n}{1 - [\kappa^{-1}n\sigma_n(\sigma_n^*)^{2h}]^{-1/2}}.$$

Note that, the use of $\widehat{A}_n^*$ instead of $\widehat{A}_n$ also improves the estimate of the asymptotically optimal bandwidth.

4 Some Monte Carlo results

To investigate the finite sample properties of our estimator by a Monte Carlo study, the spatial binary choice SAR model is specified as

$$Y_{it}^* = \lambda_0 \sum_{j=1}^{n} w_{ij} Y_{jt}^* + X_{1it} + X_{2it}\beta_0 + \alpha_i + \epsilon_{it} \tag{19}$$

for $t = 1$ and 2, where X_{1it} is drawn from the standard normal distribution $N(0,1)$, and X_{2it} from the chi-square distribution with one degree of freedom, normalized to have zero mean and unit variance, independent of each other, $\alpha_i = \frac{1}{2}(X_{2i1} + X_{2i2}) + \eta_i$ with η_i is from $N(0,1)$ independent of other variables, and ϵ_{it} is drawn from $N(0,1)$, independent of (X_{1it}, X_{2it}), the observed dependent variable Y_{it} is generated by $Y_{it} = 1$ if $Y_{it}^* > 0$ and $Y_{it} = 0$ otherwise. When the sample size is $n = 49$, the spatial weights matrix W_n corresponds to the weights matrix for the study of crimes across 49 districts in Columbus, Ohio in Anselin (1988). For large sample sizes of $n = 490$ and $n = 980$, the corresponding spatial weights matrices are block diagonal matrices with the preceding 49×49 matrix as their diagonal blocks as in Lee (2007). These correspond to the pooling, respectively, of ten and twenty separate districts with similar neighboring structures in each district.

Given that the coefficients of X could be estimated only up to scale, we set the coefficient of X_{1it} to one such that the coefficient of X_{2it} is pointly identified. In different cases of Monte Carlo study, true parameters are $\beta_0 = 1$,and $\lambda_0 = 0.3, 0.7$, respectively. As the score function $G_n(\theta, \sigma_n)$ can have many local extrema, so it is necessary to use a global optimization method such as tunneling (Levy and Montalvo, 1985) and generalized simulated annealing (Bohachevsky, Johnson, and Stein, 1986). However, results reported here are based on grid search, and the number of grid is 200.

Table 1-3 report results for comparing the performance of the estimators discussed in this paper: the spatial maximum score (SMS), smoothed spatial maximum score (SSMS), and spatial nonlinear least square (SNLLS) estimators. For SSMS and SNLLS estimators, the bandwidth for each sample is selected as follows. For SSMS, a cumulative normal distribution function is used and I compare two bandwidth selection procedures. For SSMS-1, bandwidth σ_n is selected according to the procedure suggested by Horowitz (1992) and

discussed in section 3.3. For SSMS-2, bandwidth is selected by using Silverman's rule of thumb, $\sigma_n = 1.06 \cdot \hat{s} \cdot n^{-1/5}$, where $\hat{s}$ is the sample standard deviation of Y_{it}. Finally, the bandwidth selection for SNLLS estimator is also according to Silverman's rule of thumb.

The number of repetitions is 1000 for each case in this Monte Carlo experiment. The regressors are randomly redrawn for each repetition. In each case, we report the mean bias, median bias, root mean square errors (RMSE), and mean absolute deviation (MAD) of the empirical distributions of the estimates.

Table 1 report simulation results under homoskedasticity, where ϵ_{it} is drawn from $N(0,1)$. In all the cases, the SMS and SSMS estimators perfom much better than the SNLLS estimator, this is intuitive given that the SNLLS estimator has a slower convergence rate and the bandwidth selection procedure may not be optimal. However, it is quite surprising that the SNLLS estimator seems to be biased for estimating spatial effect λ and this bias becomes larger when sample size increases, although it has a smaller bias than the SSMS estimators for the estimation of β in some cases. The performances of SSMS-1 and SSMS-2 estimators are almost the same, which suggests that Silverman's rule of thumb is an effective bandwidth selection procedure for the SSMS estimator. The SMS estimator outperfoms the SSMS estimators, especially for small sample size ($n = 49$), where the ratios of RMSEs of SMS estimator to those of the SSMS estimators are roughly 60% and 30% for the estimates of λ and β, respectively. However, this ratio increases as the n increases, which is consistent with the findings in Horowitz (1992) for the (smoothed) maximum score estimators without spatial effect. Finally, when the true parameter $\lambda_0 = 0.7$, the RMSEs of SSMS estimators for the estimation of λ is even around 30% less than those of the SMS estimator for the modest and large sample sizes.

To see the estimators are robust under heteroskedasticity and spatial errors. Table 2 reports simulation results under heteroskedasticity, where $\epsilon_{it} = \left(1 + 2Z_{it}^2 + Z_{it}^4\right) u_{it}/4$, $Z_{it} = X_{1it} + X_{2it}$, and u_{it} is logistic with median 0 and variance 1. Table 3 reports simulation results with spatial errors, where $\epsilon_{nt} = \rho_0 W_n \epsilon_{nt} + \upsilon_{nt}$, υ_{it} are i.i.d. errors with distribution $N(0,1)$, $\rho_0 = 0.5$, and weight matrix W_n is the same as in spatial lags. As we can see, the SMS and SSMS estimators appear to perform better in the heteroskedasticity designs while the SNLLS estimator appears perform worse, and all the estimators are robust under spatial errors. When $\lambda_0 = 0.7$, the SSMS estimators even outperform the

SMS estimator for the estimation of β in large sample size with spatial errors.

In summary, the SNLLS estimator has a large bias for the estimation of λ, and performs worse under heteroskedasticity. Both the SMS and the SSMS estimators delivers a robust performance for various spatial autoregressive binary choice models. The SSMS etimator can improve substantially and outperform the SMS estimator with large sample sizes.

5 Conclusion

In this paper, new estimation procedures for spatial autoregressive binary choice panel models were proposed. The estimators were based on a modification of the (smoothed) maximum score estimator to the fixed effects binary choice models without spatial effect. Asymptotic properties of the SSMS estimator were derived. A simulation study indicates these estimators perform quite well for various spatial models in finite samples.

The work here suggests areas for future research. Although both SMS and SSMS estimators have desirable asymptotic properties and perform adquately well in finite samples, they may not be easy to implement in practice. The SMS estimator has a discontinuous objective function, ruling out gradient-based optimization methods. The objective function of SSMS estimator can have several local maxima, and thus requires a global maximization algorithm that is not available in standard econometric software packages. The SNLLS estimator may be an alternative in applications, although it has a slower rate of convergence, non-Gaussian limiting distribution (Blevins and Khan, 2013), and relatively worse finite sample performance. Therefore, bias correction procedures for the SNLLS estimator or deriving other competing estimators may be the direction for future research.

Furthermore, it would be useful to explore a more effective bandwidth selection procedure than that suggested in (Horowitz, 1992), as the SSMS, especially for the bias-corrected SSMS, estimators are quite sensitive to the choice of bandwidth in finite samples.

Finally, Horowitz (2002) shows that the differences between the true and nominal levels of tests based on smoothed maximum score estimates can be very large in finite samples when first order asymptotics are used to obtain critical values, and the bootstrap provides asymptotic refinements. Thus, it is natural to ask whether this property carries on for SSMS estimator or not.

Table 1: Simulation results with homoskedasticity

	λ				β			
	Mean B	Med. B	RMSE	MAD	Mean B	Med. B	RMSE	MAD
$N = 49, \lambda_0 = 0.3$								
SMS	0.0107	-0.0023	0.0711	0.2058	-0.0009	0.0023	0.0528	0.1835
SSMS-1	0.0136	0.0339	0.1270	0.3014	0.0501	0.0791	0.1852	0.3381
SSMS-2	0.0196	0.0430	0.1313	0.2985	0.0388	0.0565	0.1758	0.3409
SNNLS	0.1199	0.4500	0.2862	0.3853	0.0370	0.0972	0.1953	0.3761
$N = 490, \lambda_0 = 0.3$								
SMS	-0.0064	0.0023	0.0455	0.1632	0.0065	-0.0068	0.0700	0.2165
SSMS-1	0.0095	0.0204	0.0455	0.1520	0.0449	0.0294	0.1176	0.2287
SSMS-2	0.0018	0.0158	0.0401	0.1579	0.0422	0.0294	0.1149	0.2289
SNNLS	0.2274	0.4500	0.3695	0.3234	0.0682	0.2668	0.2315	0.3842
$N = 980, \lambda_0 = 0.3$								
SMS	-0.0010	0.0113	0.0266	0.1305	0.0089	0.0045	0.0633	0.1951
SSMS-1	0.0105	0.0249	0.0334	0.1224	0.0274	0.0113	0.0843	0.1955
SSMS-2	0.0112	0.0249	0.0343	0.1229	0.0263	0.0113	0.0833	0.1957
SNNLS	0.2938	0.4500	0.4069	0.2538	0.1043	0.4410	0.2702	0.3868
$N = 49, \lambda_0 = 0.7$								
SMS	-0.0161	-0.0252	0.0433	0.1386	0.0021	0.0113	0.0538	0.1818
SSMS-1	0.0224	0.0842	0.0681	0.1830	0.0800	0.1786	0.2116	0.3321
SSMS-2	0.0206	0.0777	0.0657	0.1817	0.0694	0.1538	0.2005	0.3315
SNNLS	0.0656	0.2771	0.1468	0.2651	-0.0054	-0.0611	0.1379	0.3275
$N = 490, \lambda_0 = 0.7$								
SMS	-0.0140	-0.0059	0.0273	0.0931	0.0130	0.0090	0.0749	0.2102
SSMS-1	0.0086	0.0198	0.0199	0.0846	0.0443	0.0271	0.1145	0.2253
SSMS-2	0.0034	0.0134	0.0158	0.0895	0.0305	0.0204	0.1025	0.2287
SNNLS	0.2467	0.2900	0.2707	0.0766	-0.0203	-0.0814	0.1728	0.3697
$N = 980, \lambda_0 = 0.7$								
SMS	-0.0105	-0.0006	0.0195	0.0752	0.0193	0.0158	0.0717	0.1897
SSMS-1	0.0073	0.0134	0.0140	0.0649	0.0386	0.0226	0.0909	0.1898
SSMS-2	0.0021	0.0102	0.0098	0.0691	0.0339	0.0249	0.0893	0.1946
SNNLS	0.2739	0.2900	0.2834	0.0302	-0.0380	-0.1153	0.1783	0.3485

Table 2: Simulation results with heteroskedascity

	λ				β			
	Mean B	Med. B	RMSE	MAD	Mean B	Med. B	RMSE	MAD
				$N = 49, \lambda_0 = 0.3$				
SMS	-0.0043	-0.0023	0.0243	0.0978	-0.0014	-0.0023	0.0156	0.0722
SSMS-1	0.0710	0.1153	0.1593	0.2523	0.1116	0.2013	0.2287	0.3090
SSMS-2	0.0685	0.1108	0.1574	0.2540	0.1047	0.1945	0.2242	0.3127
SNNLS	0.1081	0.4410	0.2757	0.3864	0.0501	0.1696	0.2113	0.3802
				$N = 490, \lambda_0 = 0.3$				
SMS	-0.0212	-0.0158	0.0338	0.0873	0.0150	0.0181	0.0368	0.1175
SSMS-1	0.0017	0.0090	0.0147	0.0891	0.0354	0.0226	0.0671	0.1390
SSMS-2	0.0013	0.0068	0.0146	0.0903	0.0347	0.0204	0.0669	0.1403
SNNLS	0.3001	0.4500	0.4096	0.2457	0.1190	0.4274	0.2747	0.3679
				$N = 980, \lambda_0 = 0.3$				
SMS	-0.0141	-0.0090	0.0216	0.0663	0.0171	0.0113	0.0333	0.0984
SSMS-1	-0.0021	0.0023	0.0087	0.0632	0.0205	0.0158	0.0350	0.0927
SSMS-2	-0.0025	0.0023	0.0091	0.0629	0.0197	0.0113	0.0338	0.0919
SNNLS	0.3386	0.4500	0.4256	0.1938	0.1674	0.4500	0.3241	0.3620
				$N = 49, \lambda_0 = 0.7$				
SMS	-0.0088	-0.0316	0.0268	0.1088	-0.0008	0.0023	0.0295	0.1209
SSMS-1	0.0598	0.1067	0.0948	0.1520	0.0748	0.1312	0.1944	0.3094
SSMS-2	0.0578	0.1067	0.0927	0.1518	0.0659	0.1108	0.1835	0.3050
SNNLS	0.0539	0.2739	0.1372	0.2714	0.0119	-0.0430	0.1477	0.3332
				$N = 490, \lambda_0 = 0.7$				
SMS	-0.0222	-0.0155	0.0308	0.0716	0.0143	0.0204	0.0525	0.1591
SSMS-1	-0.0002	0.0102	0.0072	0.0645	0.0362	0.0158	0.0794	0.1687
SSMS-2	-0.0057	0.0038	0.0136	0.0686	0.0343	0.0158	0.0808	0.1744
SNNLS	0.2544	0.2900	0.2744	0.0637	0.0061	0.0136	0.1583	0.3690
				$N = 980, \lambda_0 = 0.7$				
SMS	-0.0150	-0.0091	0.0193	0.0517	0.0064	0.0023	0.0342	0.1322
SSMS-1	-0.0007	0.0038	0.0038	0.0440	0.0137	-0.0023	0.0377	0.1204
SSMS-2	-0.0026	0.0038	0.0062	0.0469	0.0087	-0.0023	0.0343	0.1244
SNNLS	0.2799	0.2900	0.2859	0.0191	-0.0081	-0.0226	0.1567	0.3630

Table 3: Simulation results with spatial errors

	λ				β			
	Mean B	Med. B	RMSE	MAD	Mean B	Med. B	RMSE	MAD
				$N = 49, \lambda_0 = 0.3$				
SMS	0.0071	-0.0023	0.0703	0.2130	-0.0005	0.0023	0.0579	0.1970
SSMS-1	0.0199	0.0701	0.1375	0.3091	0.0633	0.1244	0.2033	0.3458
SSMS-2	0.0216	0.0746	0.1380	0.3007	0.0551	0.0972	0.1951	0.3459
SNNLS	0.1318	0.4364	0.2883	0.3644	0.0045	0.0023	0.1656	0.3796
				$N = 490, \lambda_0 = 0.3$				
SMS	-0.0206	0.0023	0.0684	0.1837	0.0035	-0.0045	0.0740	0.2293
SSMS-1	-0.0103	0.0068	0.0600	0.1830	0.0437	0.0430	0.1253	0.2457
SSMS-2	-0.0129	0.0113	0.0640	0.1867	0.0409	0.0430	0.1232	0.2474
SNNLS	0.2189	0.4500	0.3621	0.3271	0.0641	0.2781	0.2283	0.3862
				$N = 980, \lambda_0 = 0.3$				
SMS	-0.0099	0.0023	0.0463	0.1545	-0.0029	-0.0204	0.0642	0.2079
SSMS-1	0.0042	0.0204	0.0379	0.1483	0.0188	0.0000	0.0843	0.2141
SSMS-2	0.0032	0.0204	0.0373	0.1490	0.0178	-0.0023	0.0846	0.2163
SNNLS	0.2939	0.4500	0.4061	0.2522	0.1181	0.4251	0.2769	0.3728
				$N = 49, \lambda_0 = 0.7$				
SMS	0.0103	0.0295	0.0483	0.1695	-0.0109	-0.0068	0.0726	0.2076
SSMS-1	0.0427	0.1195	0.0947	0.1970	0.0447	0.0611	0.1811	0.3378
SSMS-2	0.0389	0.1131	0.0909	0.1972	0.0342	0.0430	0.1698	0.3367
SNNLS	0.0916	0.2804	0.1663	0.2459	0.0023	-0.0430	0.1291	0.3161
				$N = 490, \lambda_0 = 0.7$				
SMS	-0.0062	0.0198	0.0290	0.1258	-0.0023	-0.0158	0.0805	0.2435
SSMS-1	0.0164	0.0424	0.0389	0.1207	0.0236	0.0158	0.1181	0.2682
SSMS-2	0.0130	0.0424	0.0360	0.1225	0.0205	0.0158	0.1136	0.2655
SNNLS	0.2415	0.2900	0.2668	0.0835	0.0059	-0.0113	0.1551	0.3646
				$N = 980, \lambda_0 = 0.7$				
SMS	-0.0137	0.0070	0.0294	0.1014	-0.0180	-0.0384	0.0850	0.2212
SSMS-1	0.0082	0.0231	0.0219	0.0932	0.0080	-0.0068	0.0836	0.2328
SSMS-2	0.0022	0.0198	0.0169	0.0970	-0.0005	-0.0158	0.0778	0.2369
SNNLS	0.2572	0.2900	0.2756	0.0596	-0.0352	-0.1560	0.1816	0.3607

Appendix A: Notations

Let $S_n^{-1} = S_n^{-1}(\lambda_0)$.

As $\frac{\partial S^{-1}}{\partial \lambda} = -S^{-1}\frac{\partial S}{\partial \lambda}S^{-1}$, we know that

$$\frac{\partial e_i^\top S_n^{-1}(\lambda)}{\partial \lambda} = e_i^\top S_n^{-1}(\lambda) W_n S_n^{-1}(\lambda);$$

Denote $B_i = K\left(\frac{e_i^\top S_n^{-1}(\lambda)\Delta X_n \beta}{\sigma_n}\right) = K\left(\frac{z_i}{\sigma_n}\right)$ where $z_i = e_i^\top S_n^{-1}(\lambda)\Delta X_n\beta$, then

$$\frac{\partial B_i}{\partial \lambda} = K'\left(\frac{z_i}{\sigma_n}\right)\frac{e_i^\top S_n^{-1}(\lambda) W_n S_n^{-1}(\lambda)\Delta X_n \beta}{\sigma_n};$$

$$\frac{\partial B_i}{\partial \tilde{\beta}^\top} = K'\left(\frac{z_i}{\sigma_n}\right)\frac{e_i^\top S_n^{-1}(\lambda)\Delta \tilde{X}_n}{\sigma_n};$$

$$B_i^{(1)}(\theta, \sigma_n) = (\partial B_i/\partial\lambda, \partial B_i/\partial\tilde{\beta}^\top)^\top;$$

$$\frac{\partial^2 B_i}{\partial \lambda^2} = K''\left(\frac{z_i}{\sigma_n}\right)\left[\frac{e_i^\top S_n^{-1}(\lambda) W_n S_n^{-1}(\lambda)\Delta X_n \beta}{\sigma_n}\right]^2 + 2K'\left(\frac{z_i}{\sigma_n}\right)\frac{e_i^\top S_n^{-1}(\lambda) W_n S_n^{-1}(\lambda) W_n S_n^{-1}(\lambda)\Delta X_n \beta}{\sigma_n};$$

$$\frac{\partial^2 B_i}{\partial \tilde{\beta}^\top \partial \lambda} = K''\left(\frac{z_i}{\sigma_n}\right)\frac{e_i^\top S_n^{-1}(\lambda) W_n S_n^{-1}(\lambda)\Delta X_n \beta}{\sigma_n}\frac{e_i^\top S_n^{-1}(\lambda)\Delta \tilde{X}_n}{\sigma_n} + K'\left(\frac{z_i}{\sigma_n}\right)\frac{e_i^\top S_n^{-1}(\lambda) W_n S_n^{-1}(\lambda)\Delta \tilde{X}_n}{\sigma_n};$$

$$\frac{\partial^2 B_i}{\partial \tilde{\beta}^\top \partial \tilde{\beta}} = K''\left(\frac{z_i}{\sigma_n}\right)\frac{\Delta \tilde{X}_n^\top [S_n^{-1}(\lambda)]^\top e_i e_i^\top S_n^{-1}(\lambda)\Delta \tilde{X}_n}{\sigma_n^2};$$

$$B_i^{(2)}(\theta, \sigma_n) = \begin{pmatrix} \frac{\partial^2 B_i}{\partial \lambda^2} & \frac{\partial^2 B_i}{\partial \tilde{\beta}^\top \partial \lambda} \\ * & \frac{\partial^2 B_i}{\partial \tilde{\beta}^\top \partial \tilde{\beta}} \end{pmatrix}$$

Recall that

$$G_n(\theta; \sigma_n) = \frac{1}{n}\sum_{i=1}^n \Delta Y_i K\left(\frac{e_i^\top S_n^{-1}(\lambda)\Delta X_n \beta}{\sigma_n}\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{Y_{i1} \neq Y_{i2}\}\left(1 - 2\cdot \mathbb{1}\{Y_{i1} = 0, Y_{i2} = 1\}\right) B_i,$$

then we have

$$T_n(\theta, \sigma_n) = \frac{\partial G_n(\theta, \sigma_n)}{\partial \tilde{\theta}} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{Y_{i1} \neq Y_{i2}\}\left(1 - 2\cdot \mathbb{1}\{Y_{i1} = 0, Y_{i2} = 1\}\right) B_i^{(1)}(\theta, \sigma_n);$$

$$Q_n(\theta, \sigma_n) = \frac{\partial^2 G_n(\theta, \sigma_n)}{\partial \tilde{\theta}\partial \tilde{\theta}^\top} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{Y_{i1} \neq Y_{i2}\}\left(1 - 2\cdot \mathbb{1}\{Y_{i1} = 0, Y_{i2} = 1\}\right) B_i^{(2)}(\theta, \sigma_n);$$

Appendix B: Proofs

Proof of Lemma 1. Without loss of generality, let $q' = q$ and consider the case in which $\beta_{0,q} > 0$ (the case $\beta_{0,q} < 0$ is symmetric). For any $(\lambda, \beta) \in \Lambda \times R^q$, let $\tilde{\beta} = (\beta_1, \ldots, \beta_{q-1})$ and $\tilde{\beta_0} = (\beta_{0,1}, \ldots, \beta_{0,q-1})$. To show that $R(\lambda, \beta) > 0$, it is sufficient to show that, for all $(\lambda, \beta) \in \Lambda \times R^q$ with $\beta/||\beta|| \neq \beta_0/||\beta_0||$, either $\Pr\left(e_i^\top S_n^{-1}(\lambda)\Delta X_n \beta < 0 < e_i^\top S_n^{-1}\Delta X_n \beta_0\right)$ or $\Pr\left(e_i^\top S_n^{-1}\Delta X_n \beta_0 < 0 < e_i^\top S_n^{-1}(\lambda)\Delta X_n \beta\right)$ or both.

Denote $e_i^\top S_n^{-1}(\lambda) = (a_{i1}(\lambda), a_{i2}(\lambda), \ldots, a_{in}(\lambda))$ and $e_i^\top S_n^{-1} = (b_{i1}(\lambda_0), b_{i2}(\lambda_0), \ldots, b_{in}(\lambda_0))$, as the inverse matrices $S_n^{-1}(\lambda)$ and S_n^{-1} exist, so there exists at least one $a_{ij}(\lambda) \neq 0$ and one $b_{ij'}(\lambda_0) \neq 0$. Apparently, there are four possible index sets: J, J', K, K', where $a_{ij}(\lambda) \neq 0, b_{ij}(\lambda_0) = 0$ for all $j \in J$; $a_{ij'}(\lambda) = 0, b_{ij'}(\lambda_0) \neq 0$ for all $j' \in J'$; $a_{ik}(\lambda) \neq 0, b_{ik}(\lambda_0) \neq 0$ for all $k \in K$; $a_{ik'}(\lambda) = 0, b_{ik'}(\lambda_0) = 0$ for all $k' \in K'$;

It is easy to see that

$$
\begin{aligned}
&\Pr\left(e_i^\top S_n^{-1}(\lambda)\Delta X_n \beta < 0 < e_i^\top S_n^{-1}\Delta X_n \beta_0\right) \\
=&\Pr\left(\sum_{j\in J} a_{ij}(\lambda)\Delta X_j \beta + \sum_{k\in K} a_{ik}(\lambda)\Delta X_k \beta < 0 < \sum_{j'=1} b_{ij'}(\lambda_0)\Delta X_{j'}\beta_0 + \sum_{k\in K} b_{ik}(\lambda_0)\Delta X_k \beta_0\right) \\
\geq&\Pr\left(A_{i,j\in J}, B_{i,j'\in J'}, C_{i,k\in K}\right)
\end{aligned}
\tag{A.1}
$$

where

$A_{i,j} = \{a_{ij}(\lambda)\Delta X_j \beta < 0\} = \{a_{ij}(\lambda)\Delta \tilde{X}_j \tilde{\beta} + a_{ij}(\lambda)\Delta X_{j,q}\beta_q < 0\}$,

$B_{i,j'} = \{b_{ij'}(\lambda_0)\Delta X_{j'}\beta_0 > 0\} = \{b_{ij'}(\lambda_0)\Delta \tilde{X}_{j'}\tilde{\beta_0} + b_{ij'}(\lambda_0)\Delta X_{j',q}\beta_{0,q} > 0\}$,

$C_{i,k} = \{a_{ik}(\lambda)\Delta \tilde{X}_k \tilde{\beta} + a_{ik}(\lambda)\Delta X_{k,q}\beta_q < 0 < b_{ik}(\lambda_0)\Delta \tilde{X}_k \tilde{\beta_0} + b_{ik}(\lambda_0)\Delta X_{k,q}\beta_{0,q}\}$

for all j, j' and k.

In the proof of Lemma 2 in Manski (1985), three cases were considered to show the positive conditional probability of $C_{i,k}$ when regarding the different signs of β_q. In our proof, we have six different cases as we need to consider the diferent signs of $a_{ik}(\lambda)$ and $b_{ik}(\lambda_0)$: (i) Case $a_{ik}(\lambda)\beta_q < 0$ and $b_{ik}(\lambda_0) > 0$:

$$
C_{i,k} = \left[\Delta X_{k,q} > \max\left(-\Delta \tilde{X}_k \tilde{\beta}/\beta_q, -\Delta \tilde{X}_k \tilde{\beta_0}/\beta_{0,q}\right)\right];
$$

(ii) Case $a_{ik}(\lambda)\beta_q < 0$ and $b_{ik}(\lambda_0) < 0$:

$$C_{i,k} = \left[-\Delta\tilde{X}_k\tilde{\beta}/\beta_q < \Delta X_{k,q} < -\Delta\tilde{X}_k\tilde{\beta}_0/\beta_{0,q} \right];$$

(iii) Case $a_{ik}(\lambda)\beta_q = 0$ and $b_{ik}(\lambda_0) > 0$:

$$C_{i,k} = \left[a_{ik}(\lambda)\Delta\tilde{X}_k\tilde{\beta} < 0, \Delta X_{k,q} > -\Delta\tilde{X}_k\tilde{\beta}_0/\beta_{0,q} \right];$$

(iv) Case $a_{ik}(\lambda)\beta_q = 0$ and $b_{ik}(\lambda_0) < 0$:

$$C_{i,k} = \left[a_{ik}(\lambda)\Delta\tilde{X}_k\tilde{\beta} < 0, \Delta X_{k,q} < -\Delta\tilde{X}_k\tilde{\beta}_0/\beta_{0,q} \right];$$

(v) Case $a_{ik}(\lambda)\beta_q > 0$ and $b_{ik}(\lambda_0) > 0$:

$$C_{i,k} = \left[-\Delta\tilde{X}_k\tilde{\beta}_0/\beta_{0,q} < \Delta X_{k,q} < -\Delta\tilde{X}_k\tilde{\beta}/\beta_q \right];$$

(vi) Case $a_{ik}(\lambda)\beta_q > 0$ and $b_{ik}(\lambda_0) < 0$:

$$C_{i,k} = \left[\Delta X_{k,q} < \min\left(-\Delta\tilde{X}_k\tilde{\beta}/\beta_q, -\Delta\tilde{X}_k\tilde{\beta}_0/\beta_{0,q} \right) \right].$$

Under Assumption 2 and using the same argument as in the proof of Lemma 2 in Manski (1985), we know that the conditional probabilities of $A_{i,j}$, $B_{i,j'}$, and $C_{i,k}$ in case (i) and (vi) are always positive. In case (iii) and (iv), we can always find a positive constant D such that the conditional probability $\Pr\left[a_{ik}(\lambda)\Delta\tilde{X}_k\tilde{\beta} - D < 0 \right] > 0$, therefore,

$$\Pr(C'_{i,k}) = \Pr\left[a_{ik}(\lambda)\Delta\tilde{X}_k\tilde{\beta} - D < 0, \Delta X_{k,q} > -\frac{\Delta\tilde{X}_k\tilde{\beta}_0}{\beta_{0,q}} \right] > 0.$$

To make sure this adjustment does not change the original conditional probability in equation (A.1), we just need to change one element in $A_{i,j}$ to $A'_{i,j} = \{a_{ij}(\lambda)\Delta X_j\beta + D < 0\}$, and the conditional probability of $A'_{i,j}$ is still positive. Using the similar adjustment, we can always make sure that the conditional probabilities $\Pr(C'_{i,k})$ in case (ii) and (v) to be positive.

Recall that in equation (A.1),

$$\begin{aligned}
&\Pr\left(e_i^\top S_n^{-1}(\lambda)\Delta X_n\beta < 0 < e_i^\top S_n^{-1}\Delta X_n\beta_0\right) \\
\geq &\Pr\left(A'_{i,j\in J}, B_{i,j'\in J'}, C'_{i,k\in K}\right) \\
= &\Pr\left(A'_{i,j\in J}|B_{i,j'\in J'}, C'_{i,k\in K}\right)\Pr\left(B_{i,j'\in J'}|C'_{i,k\in K}\right)\Pr\left(C'_{i,k\in K}\right)
\end{aligned} \tag{A.2}$$

as we just argued that each conditional probabily in equation (A.2) is positive under Assumption 2, so we have

$$\Pr\left(e_i^\top S_n^{-1}(\lambda)\Delta X_n\beta < 0 < e_i^\top S_n^{-1}\Delta X_n\beta_0\right) > 0, \qquad R(\lambda,\beta) > 0$$

Therefore, (λ_0, β_0) is identified relative to (λ, β) except those β that are scalar multiples of β_0. □

For the proof of Theorem 1, we need Proposition 1 and the following Lemmas.

Proof of Proposition 1. To prove the NED of $\{Y_{it}\}$, we first show the NED of

$$Y_{it}^* = e_i^\top S_n^{-1}(\lambda)\left(X_{nt}\beta + \alpha_n + \epsilon_{nt}\right) = \sum_{j=1}^{n} a_{ij}(\lambda)\left(X_{jt}\beta + \alpha_j + \epsilon_{jt}\right).$$

By Assumption 5 and Theorem 14.1[5] of Davidson (1994), the process $V_{it} = X_{it}\beta + \alpha_i + \epsilon_{it}$ is strong mixing with α-mixing coefficient $\alpha(m)$. Then, by the Minkowski inequality,

$$||Y_{it}^* - E(Y_{it}^*|\Im_{i,n}(m))||_2 = \left|\left|\sum_{j,d(l_i,l_j)>m} a_{ij}(\lambda)\left(V_{jt} - E(V_{jt}|\Im_{i,n}(m))\right)\right|\right|_2 \leq d_t\nu(m), \tag{A.3}$$

where $\nu(m) = \sup_i \sum_{j,d(l_i,l_j)>m} |a_{ij}(\lambda)|$, and $d_t = 2\sup_j ||V_{jt}||_2$. Therefore, $\{Y_{it}^*\}$ is NED because $\nu(m) \to 0$ as $m \to \infty$ by equation (12).

For any $\epsilon > 0$, let $\delta_\epsilon(0)$ denote the ϵ-neighborhood of 0, then we have the following inequality for the indicator function:

$$\begin{aligned}
&|\mathbb{1}\{x_1 > 0\} - \mathbb{1}\{x_2 > 0\}| \\
\leq &\frac{|x_1 - x_2|}{\epsilon}\mathbb{1}\{x_1 \notin \delta_\epsilon(0) \quad \text{or/and} \quad x_2 \notin \delta_\epsilon(0)\} + \mathbb{1}\{x_1 \in \delta_\epsilon(0), x_2 \in \delta_\epsilon(0)\}.
\end{aligned} \tag{A.4}$$

[5]The proof can be easily adjusted for the case of spatial dependence.

Denote $B = \{Y_{it}^* \in \delta_\epsilon(0), E(Y_{it}^*|\Im_{i,n}(m) \in \delta_\epsilon(0)\}$, then we have

$$
\begin{aligned}
&||\mathbb{1}\{Y_{it}^* > 0\} - E\left(\mathbb{1}\{Y_{it}^* > 0\}|\Im_{i,n}(m)\right)||_2 \\
\leq &||\mathbb{1}\{Y_{it}^* > 0\} - \mathbb{1}\{E\left(Y_{it}^*|\Im_{i,n}(m)\right) > 0\}||_2 \\
= &\left(E\left|\mathbb{1}\{Y_{it}^* > 0\} - \mathbb{1}\{E\left(Y_{it}^*|\Im_{i,n}(m)\right) > 0\}\right|^2\right)^{1/2} \\
\leq &\left(\frac{1}{\epsilon^2}\int_{B^c}|Y_{it}^* - E(Y_{it}^*|\Im_{i,n}(m))|^2\, dP + \int_B dP\right)^{1/2} \\
\leq &\frac{1}{\epsilon}\left(\int_{B^c}|Y_{it}^* - E(Y_{it}^*|\Im_{i,n}(m))|^2\, dP\right)^{1/2} + \left(\int_B dP\right)^{1/2} \\
\leq &\frac{1}{\epsilon}||Y_{it}^* - E(Y_{it}^*|\Im_{i,n}(m))||_2 + \left(\int_B dP\right)^{1/2} \\
\leq &\frac{1}{\epsilon}d_t\nu(m) + \left(\int_B dP\right)^{1/2},
\end{aligned}
$$

where the first inequality is followed by Theorem 10.12[6] of Davidson (1994), the third line is by definition, the fourth line is by equation (A.4), the last line is followed by equation (A.3). As these two terms converge to 0 when ϵ converges to 0 at a slower rate than $\nu(m)$, so the process $\{Y_{it}\}$ is near epoch dependent.

The NED of process $\{\Delta Y_i\}$ follows from Davidson (1994) Theorem 17.8, which is also applicable under spatial dependence, and the NED of $\{\text{sgn}(e_i^\top S_n^{-1}(\lambda)\Delta X_n\beta)\}$ could be shown similarly as $\{Y_{it}\}$. □

Lemma 2. Under Assumptions 1-3, and define

$$G(\theta) = \frac{1}{n}\sum_{i=1}^n G_i(\theta) = \frac{1}{n}\sum_{i=1}^n E\left[\Delta Y_i \,\text{sgn}\{e_i^\top S_n^{-1}(\lambda)\Delta X_n\beta\}\right],$$

then $G(\theta_0) > G(\theta)$ for all $\theta = (\lambda, \beta) \in \Lambda \times R^q$, where $\beta/||\beta|| \neq \beta_0/||\beta_0||$ when $\lambda = \lambda_0$.

Proof of Lemma 2. Given Lemma 1, a similar result of Manski (1987) Lemma 3 could be easily shown that $G_i(\theta_0) > G_i(\theta)$ for all individual i. That is, θ_0 uniquely maximizes each $G_i(\theta)$, so it actually maximizes $G(\theta) = \frac{1}{n}\sum_{i=1}^n G_i(\theta)$. To prove the uniqueness of $\theta_0 = \arg\max_\theta G(\theta)$, suppose there is a $\theta' \neq \theta_0$ such that θ' maximizes $G(\theta)$, which means there exists at least one m such that $G_m(\theta') \geq G_m(\theta_0)$, this contradicts with θ_0 is a unique

[6] $\mathbb{1}\{E\left(Y_{it}^*|\Im_{i,n}(m)\right) > 0\}$ is a $\Im_{i,n}(m)$ measurable approximation to $\mathbb{1}\{Y_{it}^* > 0\}$.

maximizor of $G_m(\theta)$. Therefore, we have $\theta' = \theta_0$, and θ_0 is also the unique maximizor of $G(\theta)$. □

Lemma 3. For all $c \in \mathbb{R}$ if $(\Delta Y_i, \Delta X_i)$ is strong mixing, then

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\Delta Y_i \mathbb{1}\{e_i^\top S_n^{-1}(\lambda)\Delta X_n\beta \le c\} - E\Delta Y_i \mathbb{1}\{e_i^\top S_n^{-1}(\lambda)\Delta X_n\beta \le c\}\right)\right| \to^p 0.$$

In addition, if $\alpha(m) \le Cm^{-r}$ for positive constants C and r, then convergence is almost surely.

Proof of Lemma 3. The proof is similar to the proof of Lemma 4 in de Jong and Woutersen (2011), except that ΔY_i is a heterogenous rather than stationary strong mixing process. We also apply the generic uniform law of large numbers of the Theorem of Andrews (1987). It requires compactness of the parameter space Θ, which is assumed by Assumptions 3 and 6; the summands $q_i(w_i, \theta)$, $q_i^*(w_i, \theta) = \sup\{q_i(w_i, \theta') : \theta' \in \Theta, d(\theta, \theta') < \rho\}$ and $q_{*i}(w_i, \theta) = \inf\{q_i(w_i, \theta') : \theta' \in \Theta, d(\theta, \theta') < \rho\}$ are well-defined and satisfy a (respectively weak or strong) law of large numbers; and for all $\theta \in \Theta$,

$$\lim_{\rho\to 0}\sup_i\left|\frac{1}{n}\sum_{i=1}^{n}\left(Eq_i^*(w_i,\theta) - Eq_{*i}(w_i,\theta)\right)\right| = 0.$$

Here we show the last result, denote $q_i^1(w_i, \theta) = e_i^\top S_n^{-1}(\lambda)\Delta X_n\beta$ and $K = \sup_i \sup_{\theta^*, d(\theta,\theta^*)<\rho} \frac{\partial q_i^1(w_i,\theta)}{\partial\theta}|_{\theta=\theta^*}$, we have

$$\begin{aligned}
&\lim_{\rho\to 0}\sup_n\left|\frac{1}{n}\sum_{i=1}^{n}\left(Eq_i^*(w_i,\theta) - Eq_{*i}(w_i,\theta)\right)\right| \\
=&\lim_{\rho\to 0}\sup_n\left|\frac{1}{n}\sum_{i=1}^{n}\left(E\Delta Y_i\mathbb{1}\{\sup_{\theta':d(\theta,\theta')<\rho} q_i^1(w_i,\theta') \le c\} - E\Delta Y_i\mathbb{1}\{\inf_{\theta':d(\theta,\theta')<\rho} q_i^1(w_i,\theta') \le c\}\right)\right| \\
\le&\limsup_{K\to\infty}\limsup_{\rho\to 0}\sup_n\left|\frac{1}{n}\sum_{i=1}^{n}\left[E\Delta Y_i\left(\mathbb{1}\{q_i^1(w_i,\theta) \le c+\rho K\} - \mathbb{1}\{q_i^1(w_i,\theta') \le c-\rho K\}\right)\mathbb{1}\{|w_i| \le K\}\right]\right| \\
&+\limsup_{K\to\infty}\limsup_{\rho\to 0}\sup_n\left|\frac{1}{n}\sum_{i=1}^{n}\left[E\Delta Y_i\left(\mathbb{1}\{q_i^1(w_i,\theta) \le c+\rho K\} - \mathbb{1}\{q_i^1(w_i,\theta') \le c-\rho K\}\right)\mathbb{1}\{|w_i| > K\}\right]\right| \\
\le&\limsup_{K\to\infty}\limsup_{\rho\to 0}\sup_n\left|\frac{1}{n}\sum_{i=1}^{n}\left(\Pr\{q_i^1(w_i,\theta) \le c+\rho K\} - \Pr\{q_i^1(w_i,\theta') \le c-\rho K\}\right)\right| = 0,
\end{aligned}$$

because $\Delta X_{i,q}$ has a continuous distribution. Moreover, note that $q_i(w_i, \theta)$, $q_i^*(w_i, \theta)$ and

$q_{*i}(w_i, \theta)$ are all well-defined strong mixing random variables and satisfy a strong law of large numbers of Theorem 4 of De Jong (1995) if $\alpha(m) + \nu(m) \leq Cm^{-r}$ for some positive constants C and r. □

Lemma 4. Under Assumptions 1-6, $G_n^*(\theta) \to^p G(\theta)$ uniformly over $\theta \in \Theta$, where $G_n^*(\theta) = \frac{1}{n}\sum_{i=1}^n \Delta Y_i \operatorname{sgn}\left(e_i^\top S_n^{-1}(\lambda)\Delta X_n \beta\right)$. In addition, if $\alpha(m) \leq Cm^{-r}$ for positive constants C and r, then convergence is almost surely.

Proof of Lemma 4. By equation (8), we have

$$\begin{aligned} G_n^*(\theta) &= \frac{1}{n}\sum_{i=1}^n \Delta Y_i \operatorname{sgn}\left(e_i^\top S_n^{-1}(\lambda)\Delta X_n \beta\right) \\ &= \frac{2}{n}\sum_{i=1}^n \Delta Y_i \mathbb{1}\{e_i^\top S_n^{-1}(\lambda)\Delta X_n \beta \geq 0\} - \frac{1}{n}\sum_{i=1}^n \Delta Y_i, \end{aligned}$$

both terms satisfy a weak or strong uniform law of large numbers by Lemma 3. □

Lemma 5. $G(\theta)$ is continuous at all $\theta = (\lambda, \beta^\top)^\top$ such that $\beta_q \neq 0$.

Proof of Lemma 5. The result follows from the Theorem of Andrews (1987) and Lemma 3 □

Lemma 6. Under Assumptions 1-6, $|G_n(\theta; \sigma_n) - G_n^*(\theta)| \to^p 0$ uniformly over $\theta \in \Theta$. In addition, if $\alpha(m) \leq Cm^{-r}$ for positive constants C and r, then convergence is almost surely.

Proof of Lemma 6. As in Charlier, Melenberg, and van Soest (1995), here we actually adjusted the definition of $G_n^*(\theta)$ such that $G_n^*(\theta) = \frac{1}{n}\sum_{i=1}^n \Delta Y_i \mathbb{1}\{e_i^\top S_n^{-1}(\lambda)\Delta X_n \beta \geq 0\}$ as in equation (8).

$$\begin{aligned} |G_n(\theta; \sigma_n) - G_n^*(\theta)| &= \left| \frac{1}{n}\sum_{i=1}^n \Delta Y_i \left[\mathbb{1}\{e_i^\top S_n^{-1}(\lambda)\Delta X_n \beta \geq 0\} - K\left(\frac{e_i^\top S_n^{-1}(\lambda)\Delta X_n \beta}{\sigma_n}\right)\right]\right| \\ &\leq \frac{1}{n}\sum_{i=1}^n \left| \mathbb{1}\{e_i^\top S_n^{-1}(\lambda)\Delta X_n \beta \geq 0\} - K\left(\frac{e_i^\top S_n^{-1}(\lambda)\Delta X_n \beta}{\sigma_n}\right)\right| \end{aligned}$$

Under the uniform weak or strong law of large numbers for $\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{|e_i^\top S_n^{-1}(\lambda)\Delta X_n \beta| < c\}$ converging to $\frac{1}{n}\sum_{i=1}^n \Pr\{|e_i^\top S_n^{-1}(\lambda)\Delta X_n \beta| < c\}$ (implied by Lemma 3), similar to

Horowitze (1992 Lemma 4), we can easily show that $|G_n(\theta;\sigma_n) - G_n^*(\theta)| \to 0$ almost surely uniformly over $\theta \in \Theta$ as $n \to \infty$. □

Proof of Theorem 1. For weak or strong consistency of $(\theta_n \to \theta_0)$, it is sufficient to verify the following conditions: (i) $G(\theta)$ has a unique maximum at θ_0; (ii) The parameter space Θ is compact; (iii) $G(\theta)$ is continuous; (iv) $\{G_n(\theta)\}$ converges uniformly in probability to $G(\theta)$, and strong consistency can be obtained if this is replaced by $sup_{\theta\in\Theta}|G_n(\theta) - G(\theta)| \to^{a.s.} 0$.

Condition (i) is satisfied by Lemmas 1 and 2, condition (ii) is provided by Assumptions 3 and 6, condition (iii) is proved by Lemma 5, and condition (iv) is obtained by Lemmas 4 and 6. □

For the proofs of Theorems 2 and 3, we need the following Lemmas:

Lemma 7. Let Assumptions 1-11 and 14 hold. Then

$$(a) \quad \lim_{n\to\infty} E\left[\sigma_n^{-h} T_n(\theta_0;\sigma_n)\right] = A; \qquad (b) \quad \lim_{n\to\infty} Var\left[(n\sigma_n)^{1/2} T_n(\theta_0;\sigma_n)\right] = D.$$

Proof of Lemma 7. As we know that

$$\begin{aligned} T_n(\theta_0,\sigma_n) =& \frac{1}{n}\sum_{i=1}^{n} \Delta Y_i B_i^{(1)}(\theta_0,\sigma_n) \\ =& \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Y_{i1} \neq Y_{i2}\}\left(1 - 2\cdot\mathbb{1}\{Y_{i1}=0, Y_{i2}=1\}\right) B_i^{(1)}(\theta_0,\sigma_n), \end{aligned}$$

then

$$\begin{aligned} E_n(T) =& E[\sigma_n^{-h} T_n(\theta_0,\sigma_n)] \\ =& \sigma_n^{-h}\frac{1}{n}\sum_{i=1}^{n} \Pr\{Y_{i1} \neq Y_{i2}\} \int \left[1 - 2F_i(-z_i|z_i,\tilde{Z}_i)\right] B_i^{(1)}(\theta_0,\sigma_n) p_i(z_i|\tilde{Z}_i) dz_i dP_i(\tilde{Z}_i). \end{aligned}$$

By Assumption 1, condition (5) and the Corollary of Manski (1987), we could easily derive that Median $(Y_{i1} - Y_{i2}|\Delta X_n, Y_{i1} \neq Y_{i2}) = \text{sgn}\{z_i\}$, so we have Median $(\tilde{\epsilon_i}|\Delta X_n, Y_{i1} \neq Y_{i2}) = 0$ and $F_i(0|0,\tilde{Z}_i) = 0.5$ for almost every $\tilde{Z}_i$ and $i = 1,\ldots,n$.

The proof of part (a) is analogous to that of Lemma 5 in Horowitz (1992), the only adjustment is that we need the boundedness of matrices S_n^{-1} and $S_n^{-1}W_nS_n^{-1}$ to guarantee

the boundedness of $\tilde{B}_i$ for applying Lebesgue's dominated convergence theorem. This is immediately from Assumption 6 and footnote 4.

To prove part (b), let first denote $t_n(\theta_0, \sigma_n) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{Y_{i1} \neq Y_{i2}\} B_i^{(1)}(\theta_0, \sigma_n)$, then

$$\begin{aligned}
V_n(T) =& Var\left[(n\sigma_n)^{1/2} T_n(\theta_0, \sigma_n)\right] \\
=& n\sigma_n E\left[t_n(\theta_0, \sigma_n) t_n(\theta_0, \sigma_n)^\top\right] + o(1) \\
=& \frac{\sigma_n}{n}\sum_{i=1}^n E\left[B_i^{(1)}(\theta_0, \sigma_n) B_i^{(1)}(\theta_0, \sigma_n)^\top\right] \Pr\{Y_{i1} \neq Y_{i2}\} \\
&+ \frac{\sigma_n}{n}\sum_{i=1}^n \sum_{j\neq i} E\left[B_i^{(1)}(\theta_0, \sigma_n) B_j^{(1)}(\theta_0, \sigma_n)^\top\right] \Pr\{Y_{i1} \neq Y_{i2}\} \Pr\{Y_{j1} \neq Y_{j2}\} + o(1) \\
=& D_{n1} + D_{n2} + o(1).
\end{aligned}$$

Similar to Lemma 5 of Horowitz (1992),

$$\begin{aligned}
D_{n1} =& \frac{1}{n\sigma_n}\sum_{i=1}^n E\left[\left[K'\left(\frac{z_i}{\sigma_n}\right)\right]^2 \tilde{B}_{1,i}\tilde{B}_{1,i}^\top\right] \Pr\{Y_{i1} \neq Y_{i2}\} \\
=& \frac{1}{n\sigma_n}\sum_{i=1}^n \Pr\{Y_{i1} \neq Y_{i2}\} \int \left[K'\left(\frac{z_i}{\sigma_n}\right)\right]^2 \tilde{B}_{1,i}\tilde{B}_{1,i}^\top p_i(z_i|\tilde{Z}_i) dz_i dP_i(\tilde{Z}_i) \to \quad D
\end{aligned}$$

by Lebesgue's dominated convergence theorem and Assumptions 7- 10. Lemma 7 of de Jong and Woutersen (2011) shows that D_{n2} is asymptotically negligible. This finishes the proof of part (b).

□

Lemma 8. Let Assumptions 1-11 and 14 hold. (a) If $n\sigma_n^{2h+1} \to \infty$ as $n \to \infty$, $\sigma_n^{-h} T_n(\theta_0; \sigma_n)$ converges in probability to A. (b) If $n\sigma_n^{2h+1} \to \infty$ has a finite limit κ as $n \to \infty$, $(n\sigma_n)^{1/2} T_n(\theta_0; \sigma_n)$ converges in distribution to $MVN(\kappa^{1/2}A, D)$.

Analogously to Horowitz (1992) and de Jong and Woutersen (2011), define

$$g_i(\zeta) = \mathbb{1}\{Y_{i1} \neq Y_{i2}\}\left(2 \cdot \mathbb{1}\{Y_{i1} = 1, Y_{i2} = 0\} - 1\right) \tilde{B}_{1,i} K'\left(\frac{z_i}{\sigma_n} + \zeta^\top \tilde{B}_{1,i}\right).$$

Lemma 9. If (Y_{it}, X_{it}) is strong mixing with strong mixing sequence $\alpha(m)$, and there exsit

a sequence $m_n \geq 1$ such that

$$\sigma_n^{-3(p+q-1)}\sigma_n^{-2}n^{1/s}\alpha(m_n) + (\log(nm_n))\left(n^{1-2/s}\sigma_n^4 m_n^{-2}\right)^{-1} \to 0.$$

then

$$\sup_{\zeta}\left|\frac{1}{n\sigma_n^2}\sum_{i=1}^{n}[g_i(\zeta) - Eg_i(\zeta)]\right| \to^p 0$$

Proofs of lemmas 8 and 9. The proofs are identical to the proofs of Lemma 8 and 11 of de Jong and Woutersen (2011) except that we have a different score function. □

Lemma 10. Let Assumptions 1-15 hold, and define $\phi_n = (\tilde{\theta}_n - \tilde{\theta}_0)/\sigma_n$, where θ_n is a smoothed spatial maximum score estimator. Then $\text{plim}_{n\to\infty}\phi_n = 0$.

Proof of Lemma 10. This follows from Lemma 9 and the reasoning of Lemma 8 in Horowitz (1992). □

Lemma 11. Let Assumptions 1-15 hold. Let $\{\theta_n'\} = \{\tilde{\theta}_n', \beta_{n,q}'\}$ be any sequence in Θ such that $(\theta_n' - \theta_0)/\sigma_n \to 0$ as $n \to \infty$. Then $\text{plim}_{n\to\infty} Q_n(\theta_n'; \sigma_n) = Q$.

Proof of Lemma 11. We can separately show that the elements of $Q_n(\theta; \sigma_n)$ follow a uniformly law of larger numbers. The proof is then analogous to the proof of Lemma 13 in de Jong and Woutersen (2011), except that we have different objective functions. □

Proof of Theorem 2. The proof is identical to that of Theorem 2 in Horowitz (1992), where we need Lemmas 10 and 11 instead of Lemmas 8 and 9 in Horowitz (1992). □

Proof of Theorem 3. The proof is identical to that of Theorem 7 in de Jong and Woutersen (2011), where we need Lemmas 10 and 11 instead of Lemmas 12 and 13 in de Jong and Woutersen (2011). □

References

Amemiya, T. (1985): *Advanced econometrics*. Harvard university press.

Andrews, D. W. (1987): "Consistency in nonlinear econometric models: A generic uniform law of large numbers," *Econometrica: Journal of the Econometric Society*, pp. 1465–1471.

Anselin, L. (1988): *Spatial econometrics: methods and models*, vol. 4. Springer.

Beron, K., and W. Vijverberg (2004): "Probit in a spatial context: a Monte Carlo analysis," .

Blevins, J. R., and S. Khan (2013): "Local NLLS estimation of semi-parametric binary choice models," *The Econometrics Journal*, 16(2), 135–160.

Bohachevsky, I. O., M. E. Johnson, and M. L. Stein (1986): "Generalized simulated annealing for function optimization," *Technometrics*, 28(3), 209–217.

Bolduc, D., B. Fortin, and S. Gordon (1997): "Multinomial probit estimation of spatially interdependent choices: An empirical comparison of two new techniques," *International Regional Science Review*, 20(1-2), 77–101.

Brock, W. A., and S. N. Durlauf (2001): "Discrete choice with social interactions," *The Review of Economic Studies*, 68(2), 235–260.

Chamberlain, G. (1984): "Panel data," *Handbook of Econometrics*, 2, 1247–1318.

Chamberlain, G. (1986): "Asymptotic efficiency in semi-parametric models with censoring," *Journal of Econometrics*, 32(2), 189–218.

Charlier, E., B. Melenberg, and A. H. O. van Soest (1995): "A smoothed maximum score estimator for the binary choice panel data model with an application to labour force participation," *Statistica Neerlandica*, 49(3), 324–342.

Cliff, A. D., and J. K. Ord (1981): *Spatial processes: models & applications*, vol. 44. Pion London.

Davidson, J. (1994): *Stochastic Limit Theory: An Introduction for Econometricicans*. Oxford University Press, USA.

De Jong, R. (1995): "Laws of large numbers for dependent heterogeneous processes," *Econometric Theory*, 11, 347–347.

de Jong, R. M., and T. Woutersen (2011): "Dynamic time series binary choice," *Econometric Theory*, 27(04), 673–702.

Fleming, M. (2004): "Techniques for estimating spatially dependent discrete choice models," *Advances in spatial econometrics*, pp. 145–168.

Horn, R. A., and C. R. Johnson (1985): *Matrix Analysis*. New York: Cambridge University Press.

Horowitz, J. (1992): "A smoothed maximum score estimator for the binary response model," *Econometrica: Journal of the Econometric Society*, pp. 505–531.

Horowitz, J. L. (2002): "Bootstrap critical values for tests based on the smoothed maximum score estimator," *Journal of Econometrics*, 111(2), 141–167.

Jacobs, J., A. Samarina, P. Heijnen, and P. Elhorst (2013): "State transfers at different moments in time: A spatial probit approach," Discussion paper, University of Groningen, Research Institute SOM (Systems, Organisations and Management).

Jenish, N., and I. R. Prucha (2009): "Central limit theorems and uniform laws of large numbers for arrays of random fields," *Journal of econometrics*, 150(1), 86–98.

——— (2012): "On spatial processes and asymptotic inference under near-epoch dependence," *Journal of Econometrics*.

Kelejian, H. H., and I. R. Prucha (2007): "HAC estimation in a spatial framework," *Journal of Econometrics*, 140(1), 131–154.

Khan, S. (2012): "Distribution free estimation of heteroskedastic binary response models using Probit/Logit criterion functions," *Journal of Econometrics*.

Klier, T., and D. McMillen (2008): "Clustering of Auto Supplier Plants in the United States," *Journal of Business & Economic Statistics*, 26(4), 460–471.

Lee, L. (2004): "Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models," *Econometrica*, 72(6), 1899–1925.

Lee, L.-f. (2007): "GMM and 2SLS estimation of mixed regressive, spatial autoregressive models," *Journal of Econometrics*, 137(2), 489–514.

Lee, L.-f., J. Li, and X. Lin (2013): "Binary Choice Models with Social Network under Heterogeneous Rational Expectations," *Review of Economics and Statistics*, (0).

LeSage, J. (2000): "Bayesian estimation of limited dependent variable spatial autoregressive models," *Geographical Analysis*, 32(1), 19–35.

Levy, A., and A. Montalvo (1985): "The tunneling algorithm for the global minimization of functions," *SIAM Journal on Scientific and Statistical Computing*, 6(1), 15–29.

Manski, C. (1985): "Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator," *Journal of Econometrics*, 27(3), 313–333.

——— (1987): "Semiparametric analysis of random effects linear models from binary panel data," *Econometrica: Journal of the Econometric Society*, pp. 357–362.

McFadden, D. (1974): "Conditional Logit Analysis of Qualitative Choice Behavior," *Frontiers in Econometrics*, pp. 105–142.

McMillen, D. (1992): "Probit with spatial autocorrelation," *Journal of Regional Science*, 32(3), 335–348.

Pace, R., and J. LeSage (2011): "Fast Simulated Maximum Likelihood Estimation of the Spatial Probit Model Capable of Handling Large Samples," *Available at SSRN 1966039*.

Qu, X., and L. Lee (2011): "LM tests for spatial correlation in spatial models with limited dependent variables," *Regional Science and Urban Economics*.

Wang, W., and L. Lee (2012): "Estimation of Spatial Autoregressive Models with Randomly Missing Data in the Dependent Variable," *The Econometrics Journal*.